Modelação da Taxa de Incidência de Tuberculose nas Áreas Metropolitanas de Lisboa e Porto

Sara Sofia Conceição Cerqueira
Engenharia Matemática
Departamento de Matemática
2017

Orientador
Prof. Óscar Felgueiras, Professor Auxiliar, Faculdade de Ciências da Universidade do Porto (FCUP)

Coorientador
Prof. Ana Rita Gaio, Professor Auxiliar, Faculdade de Ciências da Universidade do Porto (FCUP)

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Ao meu irmão.

Agradecimentos

Todo este percurso não poderia ter sido feito sozinha e, como tal, gostaria de agradecer a todos aqueles que fizeram parte dele.

Assim sendo, o meu agradecimento especial ao Professor Óscar Felgueiras e à Professora Ana Rita Gaio, por me terem orientado nesta dissertação, mostrando sempre disponibilidade e dedicação. Um enorme obrigada a ambos por toda a paciência, conselhos e atenção!

Um agradecimento à Professora Raquel Duarte, por ter aceite este desafio, por ter cedido os dados e por ter participado nas discussões que decorreram ao longo do ano, dando sempre uma palavra de incentivo.

Ao Daniel que, apesar de estar longe fisicamente, sempre se fez sentir por perto, ouvindo sempre os meus lamentos. Pelo apoio incondicional, preocupação e amor.

À Bárbara, por fazer sempre um esforço para estarmos juntas, mesmo quando o tempo era escasso. Obrigada por todos os conselhos e força que me deste quando mais precisei.

Por último, gostaria de agradecer à minha família, pelos ensinamentos que fizeram de mim aquilo que sou hoje. Gostaria de agradecer especialmente ao meu irmão, por todas as suas parvoíces que melhoraram os meus dias.

Resumo

A tuberculose (TB) é uma das 10 principais causas de morte em todo mundo. Nos países da União Europeia (UE) verifica-se um maior número de novos casos de TB nos grandes centros urbanos, em grupos populacionais vulneráveis. Portugal, um dos países da UE com uma das maiores taxas de incidência de TB (19.2 casos/100000 habitantes em 2015), reproduz este padrão desigual de distribuição de incidência de TB, com maior incidência nas áreas metropolitanas de Lisboa (AML) e Porto (AMP). O objectivo deste estudo foi identificar os fatores de risco (FR) que mais contribuem para a disseminação da doença nas AML e AMP, e averiguar a existência de um padrão heterogéneo entre os seus municípios. Realizou-se um estudo retrospetivo (2010-2014) nas duas áreas metropolitanas, com dados oficiais do Sistema de Vigilância de Tuberculose, Instituto Nacional de Saúde Dr. Ricardo Jorge, e Instituto Nacional de Estatística.

Dada a estrutura longitudinal dos dados, e para ter em consideração a variação regional verificada em cada área metropolitana, usou-se um modelo de regressão linear com efeitos mistos. O efeito aleatório foi identificado apenas na constante, para o único nível de agrupamento do município. Para a identificação do melhor modelo, foram consideradas diferentes estruturas fixas e aleatórias, assim como diferentes matrizes de correlação e variância residuais. A comparação entre modelos baseou-se em testes de razão de verosimilhanças para modelos aninhados e nos critérios de informação de Akaike caso contrário.

Os resultados revelaram uma associação positiva e estatisticamente significativa entre a taxa de incidência de TB e a taxa de atribuição de Rendimento Social de Inserção (RSI), a taxa de população estrangeira residente (PE) e a taxa de diagnóstico de infecção por VIH. Após o ajustamento, o modelo previu que a AMP tem uma taxa de incidência de TB maior do que a AML. Os efeitos dos fatores de risco foram os mesmos para ambas as áreas, uma vez que não foram identificados termos de interação significativos. O RSI foi o fator com maior impacto no desenvolvimento da doença. O efeito aleatório explicou 17 % da variabilidade total dos dados.

Para além da diferença na taxa de incidência de TB entre as duas áreas metropolitanas, o modelo também evidenciou uma grande variação entre os municípios da mesma região metropolitana. Os resultados obtidos sugerem a existência de uma dimensão histórico-social que influencia a incidência de TB a nível regional e que deve ser considerada quando se pensa em estratégias locais de prevenção.

Palavras-chave: Tuberculose, Dados Longitudinais, Modelo Linear Misto, Efeitos Aleatórios

Abstract

Tuberculosis (TB) is one of the 10 main causes of death in the world. In European Union (UE) countries it is observed that a higher number of new cases of TB occur in large urban areas among vulnerable population groups of residents. Portugal, one of the UE countries with one of the largest TB incidence rate (19.2 cases/100000 inhabitants in 2015), clearly demonstrates the inequality pattern of the distribution of TB incidence, with a higher incidence in the metropolitan areas of Lisbon (AML) and Porto (AMP). The objective of this study was to identify the risk factors that contribute the most to the spread of TB in AML and AMP, and whether there is a pattern of heterogeneous TB incidence among its municipalities. A retrospective study (2010-2014) was carried out, with official data from the Tuberculosis Surveillance System, National Institute of Health Dr. Ricardo Jorge and National Statistical Institute, in the two metropolitan areas.

The intra-individual (municipalities) variability and the longitudinal data profile were taken into account by the fitting of a linear mixed-effects regression model. The random effect was identified at the intercept, only, for the single municipality level of grouping. For the identification of the best model, regressions with different fixed linear predictors, random effects structures, residual correlation matrixes and residual variances were considered. Comparison between models was based on the likelihood ratio test for nested models and on the Akaike Information Criteria otherwise.

TB incidence was positively (and significantly) associated with beneficiaries of social income (RSI), foreign population (FP) and infection by the human immunodeficiency virus (VIH). After adjustment for these covariates, the model predicted AMP to have a larger TB incidence rate than AML. The effects of the risk factors were the same for both areas, as no significant interaction terms were identified. RSI was the most influential variable on TB incidence. The random effect in the model accounted for 17% of the total data variability.

Besides the difference in TB incidence rate between the two metropolitan areas, the model has also evidenced a large variation between municipalities from the same metropolitan area. The results suggest the existence of a social-historical dimension that influences TB incidence at the regional level and that should be considered when thinking about local prevention strategies.

Keywords: Tuberculosis, Longitudinal data, Linear Mixed-Effects Model, Random Effects

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AIC Critério de Informação de Akaike.

AML Área Metropolitana de Lisboa.

AMP Área Metropolitana do Porto.

BCG Bacilo de Calmette e Guérin.

BD Base de Dados.

BIC Critério de Informação Bayesiana.

BLUP Melhor Preditor Linear Centrado.

CI Código de Identificação.

DGS Direção-Geral da Saúde.

EBLUP Melhor Preditor Linear Centrado Empírico.

EM Estimação-Maximização.

FR Fatores de Risco.

GL Graus de Liberdade.

INE Instatuto Nacional de Estatística.

INSA Instituto Nacional de Saúde Dr. Ricardo Jorge.

MLG Modelo Linear Misto Geral.

MLM Modelo Linear Misto.

MLMB Modelo Linear Misto Básico.

MV Máxima Verosimilhança.

MVR Máxima Verosimilhança Restrita.

NR Newton-Raphson.

NUTS 3 Nomenclatura das Unidades Territoriais para Fins Estatísticos, Nível 3.

OMS Organização Mundial da Saúde.

QQ Quantil-quantil.

RSI Rendimento Social de Inserção.

SD Desvio Padrão.

SE Erro Padrão.

SVIG-TB Sistema de Vigilância da Tuberculose.

TB Tuberculose.

TRV Teste de Razão de Verosimilhanças.

UE União Europeia.

VIH Vírus da Imunodeficiência Humana.

Capítulo 1

Introdução

1.1 Motivação

A tuberculose (TB) é uma doença com que a humanidade lida há milhares de anos, marcando a sua presença ao longo da história: foram encontradas deformidades típicas de TB em ossadas mumificadas no Egito, textos escritos há mais de 2 mil anos descrevem a doença na China e na Índia, assim como na Grécia Antiga, onde a doença era denominada *phthisis*.

Na Idade Média, a TB proliferou progressivamente, ultrapassando a lepra, e atingindo o pico entre os séculos XVIII e XIX, altura em que os trabalhadores das áreas rurais se deslocavam para as cidades à procura de trabalho. Nalgumas partes da Europa era conhecida como o "Mal do Rei", e existia a crença de que podia ser curada pelo toque do monarca de Inglaterra ou de França. Caricaturas como esta eram frequentemente publicadas, tanto como distrações humorísticas, como ilustrações da natureza séria da doença:

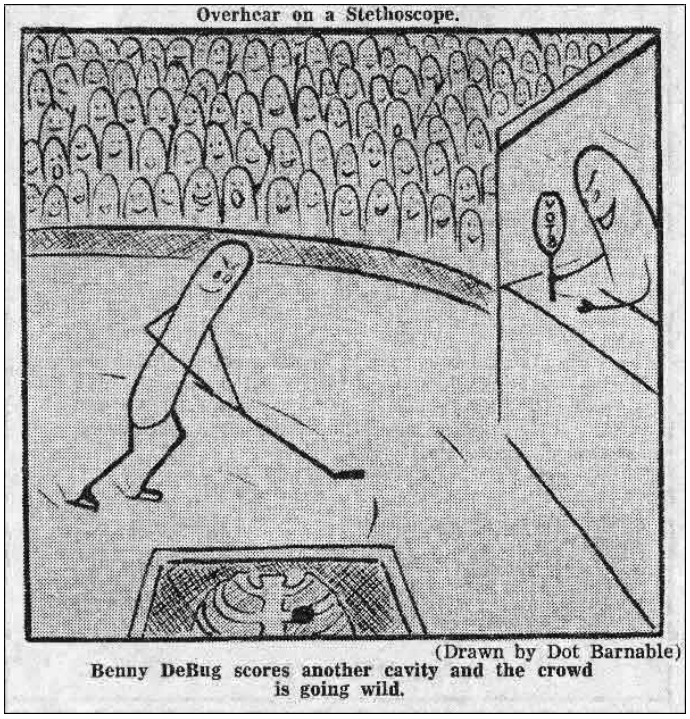


Figura 1.1: Caricatura do bacilo da TB

Apesar de Robert Koch ter descoberto o micro-organismo responsável pela TB a 24 de Março de 1882[1], foi somente no século posterior, em 1913, que foi descoberto um micro-organismo capaz de resistir e prevenir a doença, o Bacilo de Calmette e Guérin (BCG). Os retratos dos autores responsáveis pela descoberta de tal micro-organismo podem ser visualizados na figura 1.2 (b) e (c). Contudo, a cura da TB apenas se tornou possível a partir de 1944, ano em que foi descoberta a estreptomicina. Esta substância, associada a antibacilares, permitiu a cura de quase todos os casos. Na figura 1.2 (a) podemos visualizar o retrato de Robert Koch, enquanto que em (b) e (c) podemos visualizar os retratos dos autores responsáveis pela descoberta do BCG, Albert Calmette e Camille Guérin, respetivamente.

(a) Robert Koch

(b) Albert Calmette

(c) Camille Guérin

Figura 1.2: Autores dos avanços médicos que levaram à cura da TB

A diminuição constante do número de casos de TB ao longo dos anos 60 e 70 levou especialistas de saúde pública a prever a quase erradicação da doença até ao ano 2000. No entanto, essa tendência interrompeu-se abruptamente em meados dos anos 80 com o surgimento de uma nova epidemia: o vírus da imunodeficiência humana (VIH). Em vez de desencadear uma resposta imunológica, o VIH destrói insidiosamente os próprios mecanismos do corpo para o combate de infeções. Por este motivo, o cruzamento destas duas doenças contagiosas levou ao reaparecimento das formas mais graves de TB e à ineficiência dos fármacos usados no seu tratamento. Perante esta situação, em Abril de 1993, a Organização Mundial da Saúde (OMS) declarou a doença como emergência mundial.

Os avanços científicos e tecnológicos nesta área, assim como o desenvolvimento de uma vacina e de um tratamento eficaz, tornaram possível controlar a doença. No entanto, a TB ainda é considerada uma das dez principais causas de morte em todo o mundo sendo que, segundo o último relatório global da OMS, em 2015, surgiram aproximadamente 10.4 milhões de novos casos e houve cerca de 1.8 milhões de mortes causadas por esta doença. [44] Nesse mesmo ano, em Portugal, foram diagnosticados e comunicados à Direção-Geral da Saúde (DGS) 2158 casos de TB. Destes, apenas 1987 eram novos casos, traduzindo assim uma taxa de incidência de 19.2 por 100 mil habitantes. Esta é uma situação particularmente grave se pensarmos que se trata de uma doença curável, cujo tratamento não é dispendioso. Assim sendo, existe a necessidade de realizar estudos que nos permitam averiguar quais os fatores de risco que levam à proliferação desta doença. [12]

No relatório da DGS sobre o Programa Nacional para a Infeção VIH/SIDA e Tuberculose consta a figura 1.3, onde podemos observar o número de casos de TB por distrito no ano 2015 e verificar que existe uma grande assimetria a nível nacional: os distritos de Lisboa e do Porto

[1]Data que passou a ser assinalada como Dia Mundial da Tuberculose.

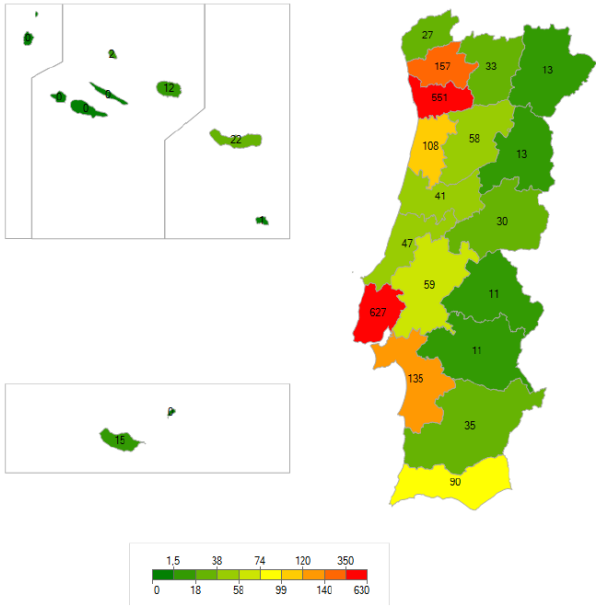


Figura 1.3: Número de casos de TB por distrito no ano de 2015

são os que apresentam os valores mais altos de incidência absoluta, havendo ainda uma enorme discrepância relativamente aos restantes distritos.

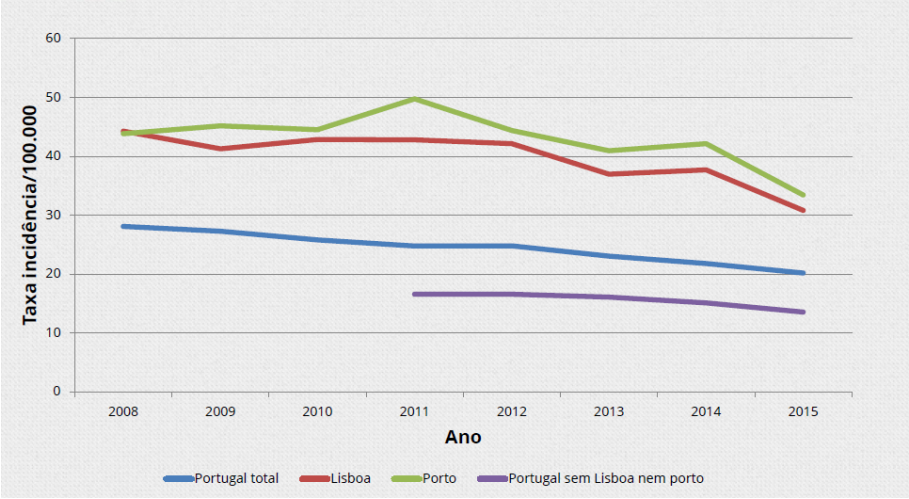


Figura 1.4: Discrepâncias na taxa de incidência de TB ao longo dos anos [13]

É da análise do gráfico na figura 1.4, do mesmo relatório, que surge a motivação final para esta dissertação: os distritos de Lisboa e do Porto têm um grande impacto na taxa de incidência a nível nacional, sendo que, sem contar com estes, a taxa de incidência diminui drasticamente.

1.2 Objetivos

Nos países da União Europeia, verifica-se que os maiores números de novos casos de TB ocorrem em grandes áreas urbanas, entre grupos populacionais vulneráveis: emigrantes, sem-abrigo, pessoas com histórico de abuso de drogas e álcool, assim como pessoas com co-infecção por VIH. Outros fatores de risco (FR) associados a grandes cidades são condições precárias de

habitação (como espaços lotados e mal ventilados) e difícil acesso a serviços de saúde, o que pode levar a diagnósticos e tratamento tardios. [45, 41, 5]

Sendo assim, o objectivo principal deste estudo é analisar e comparar a incidência de TB nas duas grandes áreas metropolitanas do país, tendo em conta um grupo de fatores de risco pré-definidos. O foco do estudo foi tentar identificar quais os FR que mais contribuem para a disseminação da doença em cada uma das regiões e perceber se há um padrão de comportamento heterogéneo entre elas. Desta forma, pretende-se que a divulgação dos resultados obtidos seja útil na construção de medidas de combate e prevenção da doença direcionadas para estes grupos de risco, podendo assim contribuir para a diminuição da taxa de incidência nestas regiões e, consequentemente, a nível nacional.

A escolha de incidir este estudo sobre as Áreas Metropolitanas de Lisboa e do Porto, em vez dos respetivos distritos, provém de um estudo divulgado pela DGS (que consta no mesmo relatório mencionado na secção anterior). Neste, foram detetados clusters dos municípios com maior e menor taxa de incidência e, como podemos verificar pela figura 1.5 (resultado do estudo referido), os municípios que pertencem aos clusters de maior incidência ultrapassam os limites geográficos dos distritos de Lisboa e do Porto.

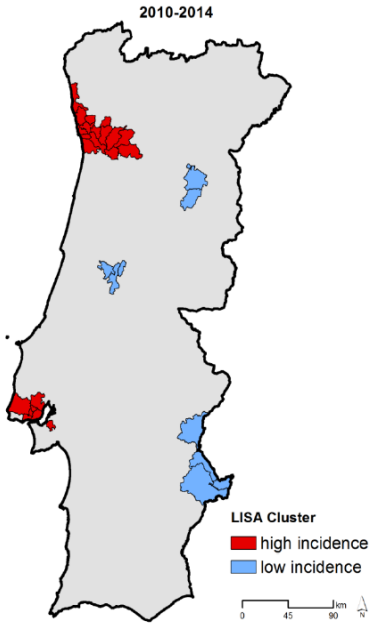


Figura 1.5: Deteção de clusters de maior e menor taxa de incidência de TB

De modo a poder contemplar algumas destas regiões, optou-se por focar o estudo nas Áreas Metropolitanas de Lisboa e Porto, definidas pelo nível 3 da Nomenclatura das Unidades Territoriais para Fins Estatísticos (NUTS 3). Outro estudo recente, no qual foram analisados todos os casos de TB notificados em Portugal de 2010 a 2014, também revelou que os municípios de maior risco de incidência se encontram nestas duas áreas metropolitanas. [2]

Na tabela 1.1 pindicamos os concelhos de cada uma das regiões, assim como os respetivos códigos de identificação (CI).

O facto de haver concelhos pertencentes à mesma área metropolitana com taxas de incidência muito díspares, apesar da proximidade geográfica e da semelhança em termos de contexto socio-

Tabela 1.1: Concelhos e respetivos códigos de identificação da AML e da AMP

AML		**AMP**	
CI	Concelho	CI	Concelho
1	Alcochete	19	Arouca
2	Almada	20	Espinho
3	Amadora	21	Gondomar
4	Barreiro	22	Maia
5	Cascais	23	Matosinhos
6	Lisboa	24	Oliveira de Azeméis
7	Loures	25	Paredes
8	Mafra	26	Porto
9	Moita	27	Póvoa de Varzim
10	Montijo	28	Santa Maria da Feira
11	Odivelas	29	Santo Tirso
12	Oeiras	30	São João da Madeira
13	Palmela	31	Trofa
14	Seixal	32	Vale de Cambra
15	Sesimbra	33	Valongo
16	Setúbal	34	Vila do Conde
17	Sintra	35	Vila Nova de Gaia
18	Vila Franca de Xira		

económico das populações, coloca a questão de poder haver variação regional no que se refere ao peso que cada fator de risco tem no desenvolvimento da doença. Esta constatação sugere assim que os centros urbanos não devem ser tratados como entidades homogéneas no que se refere ao estudo da TB. [2, 39]

Os dados a analisar são referentes aos anos 2010, 2011, 2012, 2013 e 2014, perfazendo assim um total de 175 observações (5 para cada um dos 35 municípios considerados). É então do nosso interesse estudar e aplicar um modelo de regressão que seja capaz de analisar dados em que os indivíduos são medidos/observados ao longo do tempo (dados longitudinais). O modelo escolhido para o estudo em questão é o Modelo Linear Misto, uma vez que incorpora efeitos fixos, que são parâmetros associados à população inteira, e efeitos aleatórios, associados às unidades de análise/indivíduos e que são usados diretamente na modelação da variabilidade inter-indivíduos.

Para além da variável resposta, a taxa de incidência de TB (cedida pelo SVIG-TB[2]), os dados contemplam 26 variáveis explicativas: dados cedidos pelo Instituto Nacional de Saúde Dr. Ricardo Jorge (INSA) que traduzem o número de casos de VIH por 10^5 habitantes, e ainda 26 outras variáveis recolhidas do Instituto Nacional de Estatística (INE) que traduzem FR conhecidos e relacionados com a TB, como por exemplo, a proporção de população masculina, a densidade populacional e a proporção de médicos.

A implementação dos modelos acima referidos foi conduzida no software R (versão 3.3.2), através da biblioteca *nlme*. [34]

[2]Sistema de Vigilância da Tuberculose

1.3 Estrutura

Esta dissertação encontra-se dividida em quatro capítulos, incluindo esta introdução.

O Capítulo 2 é dedicado à revisão de literatura, onde será exposta a notação e os conceitos teóricos do Modelo Linear Misto. Neste, apresentamos a especificação geral do modelo, métodos de estimação baseados na máxima verosimilhança e na máxima verosimilhança restrita, assim como exemplos de algoritmos de otimização que estes métodos de estimação requerem, predição dos efeitos aleatórios, testes de razão de verosimilhanças e critérios de informação, diagnóstico do modelo através da verificação das condições impostas e, por fim, modelação da heterocedasticidade e da correlação residual.

O Capítulo 3 é dedicado ao estudo do caso em questão. Começamos por apresentar uma breve descrição da base de dados, seguida de uma análise exploratória. Posteriormente, são apresentados os resultados obtidos da aplicação dos modelos descritos no Capítulo 2, bem como a respetiva interpretação.

No Capítulo 4 é apresentada uma discussão dos resultados obtidos no Capítulo 3 e das limitações inerentes ao estudo que poderão ser aprimoradas em trabalhos futuros.

Capítulo 2

Modelo Linear Misto

O Modelo Linear Misto (MLM) tem como objetivo principal descrever a relação linear entre uma variável resposta contínua e algumas covariáveis, em dados que são agrupados de acordo com um ou mais fatores de classificação. Um exemplo de dados que podem ser analizados através destes modelos estatísticos são os dados longitudinais, nos quais os indivíduos são medidos repetidamente ao longo do tempo (dados obtidos de forma prospetiva) ou em condições diferentes (forma retrospetiva), em relação a uma característica, sendo o próprio tempo um fator de interesse. Medições feitas à mesma variável para o mesmo indivíduo são, provavelmente, correlacionadas, pelo que ajustar dados longitudinais envolve estimar parâmetros de covariância capazes de capturar esta correlação.

O nome Modelo Linear Misto deve-se ao facto de este incorporar **efeitos fixos**, que são parâmetros associados à população inteira, e **efeitos aleatórios**, que são variáveis aleatórias associadas às unidades de análise recolhidas aleatoriamente da população.

2.0.1 Efeitos Fixos vs. Efeitos Aleatórios

Os **efeitos fixos**, também denominados coeficientes de regressão, descrevem a relação entre a variável dependente e as variáveis preditivas, para uma população, ou para um número relativamente pequeno de subpopulações definidas por níveis de um fator fixo. Podem estar associados a covariáveis contínuas (como o peso ou a temperatura) ou categóricas (como o género ou o país de origem). Isto é, podem descrever contrastes ou diferenças entre níveis de um fator fixo (por exemplo, diferenças entre o sexo masculino e o sexo feminino) em termos da resposta média da variável contínua dependente, ou o efeito de uma covariável contínua.

Os **efeitos aleatórios** são variáveis aleatórias associadas aos aos níveis de um dado fator aleatório no MLM. Estes valores, que são específicos a um dado nível, representam os desvios das observações/unidades experimentais em relação à componente fixa do preditor linear. Por exemplo, podem representar desvios aleatórios para um dado indivíduo em relação ao termo constante (fixo) global - efeito aleatório na constante - ou em relação a um dado efeito fixo global (coeficientes aleatórios). Ao contrário dos efeitos fixos, que são representados por parâmetros constantes num MLM, os efeitos aleatórios são representados por variáveis aleatórias, sobre as quais se assume uma distribuição normal.

2.1 O Modelo

Nesta secção apresentamos formalmente o MLM no contexto específico de dados longitudinais com um único nível de agrupamento. Na exposição que se segue, o índice i é referente aos indivíduos (unidades de análise), e o índice t é indicativo do instante de tempo.

Consideremos uma amostra de N indivíduos, medida repetidamente ao longo do tempo. Por uma questão de simplicidade, começamos por formular o MLM ao nível de uma observação individual. Nesta terminologia, Y_{it} representa a medição da variável resposta contínua, $\mathbf{Y}$, observada para o i-ésimo indivíduo no t-ésimo instante de tempo. Os dados que mais à frente iremos analisar são equilibrados, isto é, não existem observações omissas: todos os indivíduos têm o mesmo número de medições, e todas as medições foram realizadas nas mesmas ocasiões. Por este motivo, ao longo deste estudo, vamos assumir que são feitas n medições da variável resposta para cada indivíduo, e que existem N indivíduos na amostra. Em Fitzmaurice *et al.* (1994) é apresentada uma discussão sobre dados não equilibrados, com exemplos concretos e implicações que diferentes tipos de mecanismos de dados omissos têm na análise.

Podemos, então, escrever a variável resposta para o indivíduo i no t-ésimo instante como:

$$Y_{it} = \underbrace{\beta_1 X_{it}^{(1)} + \beta_2 X_{it}^{(2)} + \cdots + \beta_p X_{it}^{(p)}}_{\text{Parte fixa}} + \underbrace{b_{i1} Z_{it}^{(1)} + b_{i2} Z_{it}^{(2)} + \cdots + b_{iq} Z_{it}^{(q)}}_{\text{Parte aleatória}} + e_{it},$$

onde e_{it} representa o erro associado à previsão dada pela soma da parte fixa com a parte aleatória. Assumimos que existem dois conjuntos de covariáveis, nomeadamente, os conjuntos de covariáveis $\mathbf{X}$ e $\mathbf{Z}$. O primeiro conjunto contém p covariáveis, $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)}$, associadas aos efeitos fixos $\beta_1, \ldots, \beta_p$; o segundo contém q covariáveis, $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(q)}$, associadas aos efeitos aleatórios $b_{i1}, \ldots, b_{iq}$ que são específicos para o indivíduo i. As covariáveis destes conjuntos podem ser contínuas ou categóricas.

Dada uma covariável $\mathbf{X}^{(k)}$, $k = 1, \ldots, p$, representamos o seu valor no indivíduo i e instante de tempo t por $X_{it}^{(k)}$. Assumimos que o conjunto X pode conter dois tipos de covariáveis:

- Covariáveis invariantes no tempo: características individuais do sujeito que não se alteram ao longo da duração do estudo. Um exemplo de uma covariável invariante no tempo é o sexo dos indivíduos (feminino ou masculino);

- Covariáveis variáveis no tempo: características individuais que sofrem alterações a cada medição, como por exemplo, o tempo de medição, altura dos indivíduos, etc. A inclusão de covariáveis cujos valores individuais sofrem alterações ao longo do tempo introduz dificuldades em relação à interpretação e à estimação dos modelos resultantes.

Cada parâmetro β_i representa o efeito (fixo) que a alteração de uma unidade de medida em $\mathbf{X}_i$ tem no valor esperado da variável dependente, quando todas as outras covariáveis permanecem constantes. Estes parâmetros são os efeitos fixos a estimar e a relação linear entre estes e as covariáveis de $\mathbf{X}$ definem a parte fixa do modelo.

Os efeitos das das covariáveis $\mathbf{Z}$ sobre a resposta são representados na parte aleatória do modelo pelos q efeitos aleatórios, $\mathbf{b}_1, \ldots, \mathbf{b}_q$. A premissa básica do MLM é que a heterogeneidade entre indivíduos é modelada por este subconjunto de coeficientes de regressão.

Os erros aleatórios são denotados por e_{it} e representam os desvios da variável resposta ao seu valor esperado previsto pelo preditor linear.

Na próxima secção iremos combinar as observações individuais em notação vetorial e matricial.

2.1.1 Notação vetorial e matricial

Representamos por $\mathbf{Y}_i$ o vector das respostas (contínuas) do indivíduo i. Os elementos de $\mathbf{Y}_i$ podem ser escritos usando a notação vista na secção anterior para observações individuais:

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{bmatrix}.$$

Supõe-se que existe independência inter-indivíduos. Isto é, que os vetores das respostas, $\mathbf{Y}_i$, são independentes entre si. No entanto, não se pode assumir independência entre elementos de $\mathbf{Y}_i$ (medições realizadas à mesma unidade de análise).

O MLM escreve $\mathbf{Y}_i$ na seguinte forma:

$$\begin{aligned} &\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \\ &\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}), \\ &\mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i). \end{aligned} \tag{2.1}$$

Aqui, $\mathbf{X}_i$ é uma matriz $n \times p$ que representa os valores (conhecidos) das p covariáveis, $X^{(1)}, \ldots, X^{(p)}$, para cada um dos n instantes temporais observados para o i-ésimo indivíduo:

$$\mathbf{X}_i = \begin{bmatrix} X_{i1}^{(1)} & X_{i1}^{(2)} & \ldots & X_{i1}^{(p)} \\ X_{i2}^{(1)} & X_{i2}^{(2)} & \ldots & X_{i2}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in}^{(1)} & X_{in}^{(2)} & \ldots & X_{in}^{(p)} \end{bmatrix},$$

$X_{it}^{(k)}$ corresponde ao valor observado da covariável k, no t-ésimo instante, para o i-ésimo indivíduo. É de notar que, para uma covariável invariante no tempo, todos os elementos da coluna correspodente seriam iguais. Num modelo com termo constante, a primeira coluna seria igual a 1 para todas as observações.

Assume-se que nenhuma coluna (ou linha) de $\mathbf{X}_i$ é uma combinação linear das restantes.

O parâmetro $\boldsymbol{\beta}$ na equação (2.1) é o vetor dos p coeficientes de regressão desconhecidos (efeitos fixos) associados às p covariáveis usadas na construção da matriz $\mathbf{X}_i$:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

A matriz $n \times q$, $\mathbf{Z}_i$, representa os valores conhecidos das q covariáveis dos efeitos aleatórios, $Z^{(1)}, \ldots, Z^{(q)}$, para o i-ésimo indivíduo. Esta matriz é, em geral, um subconjunto de covariáveis de $\mathbf{X}_i$.

$$\mathbf{Z}_i = \begin{bmatrix} Z_{i1}^{(1)} & Z_{i1}^{(2)} & \ldots & Z_{i1}^{(q)} \\ Z_{i2}^{(1)} & Z_{i2}^{(2)} & \ldots & Z_{i2}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{in}^{(1)} & Z_{in}^{(2)} & \ldots & Z_{in}^{(q)} \end{bmatrix}.$$

As colunas de $\mathbf{Z}_i$ representam os valores observados das q variáveis preditivas, para o i-ésimo indivíduo, cujos efeitos na resposta contínua variam aleatoriamente de acordo com o sujeito. Num MLM no qual apenas o termo constante é assumido aleatório, a matriz $\mathbf{Z}_i$ seria simplesmente uma coluna de 1's.

O vetor $\mathbf{b}_i$, para o i-ésimo indivíduo, representa o vetor dos q efeitos aleatórios associados às q covariáveis da matriz $\mathbf{Z}_i$.

$$\mathbf{b}_i = \begin{bmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{iq} \end{bmatrix}, \quad \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}).$$

Assume-se que o vetor $\mathbf{b}_i$ segue uma distribuição normal multivariada, com média (vetor) $\mathbf{0}$ e matriz de variância-covariância $\mathbf{D}$.

Os elementos diagonais da matriz $\mathbf{D}$ representam as variâncias de cada efeito aleatório em $\mathbf{b}_i$, enquanto que os restantes elementos representam as covariâncias entre cada par de efeitos aleatórios. Uma vez que existem q efeitos aleatórios associados ao indivíduo i, $\mathbf{D}$ é uma matriz $q \times q$, simétrica e definida positiva.

$$\mathbf{D} = \mathrm{Var}(\mathbf{b}_i) = \begin{bmatrix} \mathrm{Var}(b_{i1}) & \mathrm{Cov}(b_{i1}, b_{i2}) & \ldots & \mathrm{Cov}(b_{i1}, b_{iq}) \\ \mathrm{Cov}(b_{i1}, b_{i2}) & \mathrm{Var}(b_{i2}) & \ldots & \mathrm{Cov}(b_{i2}, b_{iq}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(b_{i1}, b_{iq}) & \mathrm{Cov}(b_{i2}, b_{iq}) & \ldots & \mathrm{Var}(b_{iq}) \end{bmatrix}.$$

Os elementos da matriz $\mathbf{D}$ são definidos como funções de um conjunto pequeno de parâmetros de covariância, armazenados num vetor denotado $\boldsymbol{\theta_D}$. É de notar que o vetor $\boldsymbol{\theta_D}$ impõe uma estrutura (ou restrições) aos elementos da matriz $\mathbf{D}$. Mais à frente iremos discutir as diferentes estruturas para esta matriz.

Finalmente, o vetor $\mathbf{e}_i$ é o vetor dos n resíduos, onde cada elemento representa o erro associado à resposta observada no instante t para o i-ésimo indivíduo:

$$\mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in} \end{bmatrix}, \quad \mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i).$$

Em contraste com o modelo linear simples, os resíduos associados a observações realizadas no mesmo indivíduo podem ser correlacionados. Assume-se que $\mathbf{e}_i$ é um vetor aleatório seguindo uma distribuição normal multivariada com média (vetor) $\mathbf{0}$ e matriz de variância-covariância $\boldsymbol{\Sigma}_i$. Assumimos também que os resíduos associados a diferentes indivíduos são independentes entre si.

Os resíduos, $\mathbf{e}_i$, e os efeitos aleatórios, $\mathbf{b}_j$, são supostos independentes para indivíduos diferentes.

A matriz diagonal $\boldsymbol{\Sigma}_i$, para o i-ésimo indivíduo, tem a seguinte forma:

$$\boldsymbol{\Sigma}_i = \mathrm{Var}(\mathbf{e}_i) = \sigma^2\mathbf{I} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix},$$

em que os resíduos associados a observações realizadas ao mesmo indivíduo são independentes e com igual variância (σ^2).

Com base no modelo (2.1) e nas condições impostas anteriormente, podemos concluir que o vetor das respostas, $\mathbf{Y}_i|\mathbf{X}_i$, tem distribuição normal multivariada, cujos parâmetros podem ser calculados através de:

$$\begin{aligned}
\mathrm{E}(\mathbf{Y}_i|\mathbf{X}_i) &= \mathrm{E}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i) \\
&= \mathrm{E}(\mathbf{X}_i\boldsymbol{\beta}) + \mathrm{E}(\mathbf{Z}_i\mathbf{b}_i) + \mathrm{E}(\mathbf{e}_i) \\
&= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\underbrace{\mathrm{E}(\mathbf{b}_i)}_{=0} + \underbrace{\mathrm{E}(\mathbf{e}_i)}_{=0} \\
&= \mathbf{X}_i\boldsymbol{\beta}, \\
\mathrm{Var}(\mathbf{Y}_i|\mathbf{X}_i) &= \mathrm{Var}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i) \\
&= \underbrace{\mathrm{Var}(\mathbf{X}_i\boldsymbol{\beta})}_{=0} + \mathrm{Var}(\mathbf{Z}_i\mathbf{b}_i) + \mathrm{Var}(\mathbf{e}_i) \\
&= \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \boldsymbol{\Sigma}_i.
\end{aligned}$$

Ou seja, $\mathbf{Y}_i|\mathbf{X}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \boldsymbol{\Sigma}_i)$ e tem função densidade de probabilidade (f.d.p.) dada por:

$$f(\mathbf{Y}_i|\mathbf{X}_i) = (2\pi)^{-\frac{n}{2}}|\mathbf{V}_i|^{-\frac{1}{2}}\exp\left(-\frac{(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})}{2}\right), \tag{2.2}$$

onde $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \boldsymbol{\Sigma}_i$. Outra forma de obter a f.d.p. de $\mathbf{Y}_i|\mathbf{X}_i$, seria através da seguinte relação:

$$f(\mathbf{Y}_i|\mathbf{X}_i) = \int f(\mathbf{Y}_i|\mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i,$$

com

$$f(\mathbf{Y}_i|\mathbf{b}_i) = (2\pi)^{-\frac{n}{2}} |\mathbf{\Sigma}_i|^{-\frac{1}{2}} \exp\left(-\frac{(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)^\top \mathbf{\Sigma}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)}{2} \right),$$

e

$$f(\mathbf{b}_i) = (2\pi)^{-\frac{q}{2}} |\mathbf{D}|^{-\frac{1}{2}} \exp\left(-\frac{\mathbf{b}_i^\top \mathbf{D} \mathbf{b}_i}{2} \right).$$

Para completar a notação do MLM, introduzimos nas secções seguintes o vetor $\boldsymbol{\theta}$, que combina todos os parâmetros de covariância contidos no vetor $\boldsymbol{\theta_D}$, assim como o parâmetro de variância associado à matriz $\mathbf{\Sigma}_i$ (σ^2).

2.2 Matriz de variância-covariância não estruturada

Quando o número de medições é relativamente pequeno, é razoável permitir que a matriz de variância-covariância seja arbitrária, isto é, sem qualquer restrição aos seus elementos. A única exigência imposta é que seja simétrica e definida positiva. Quando não é assumida uma estrutura explicita, a matriz resultante é referida como covariância não estruturada, e tem a vantagem de não haver suposições sobre as variâncias e as covariâncias.

Ao tentar impor alguma estrutura à matriz de variância-covariância, é necessário atingir um equilíbrio: pouca informação sobre a estrutura vai resultar numa quantidade excessiva de parâmetros a estimar dada a quantidade limitada de dados disponível. Esta é a principal desvantagem da covariância não estruturada: ao não considerar restrições aos elementos da matriz, o número de parâmetros a estimar vai crescer exponencialmente com o número de medições. É possível melhorar a precisão com que os parâmetros são estimados impondo estruturas à matriz de variância-covariância. No entanto, demasiadas restrições aos elementos desta matriz podem resultar no risco de má especificação do modelo, originando assim inferências erradas.

Nas próximas secções iremos descrever algumas das estruturas de covariância mais usadas em modelos longitudinais. Em Fitzmaurice *et al.* (1994) podemos ver outros exemplos de estruturas de covariância, que também se encontram disponíveis em vários softwares.

2.2.1 Estruturas de covariância: matriz D

Por defeito, a matriz de variância-covariância dos efeitos aleatórios é uma matriz simétrica e definida positiva não estruturada. No entanto, em muitas situações práticas, temos interesse em restringir $\mathbf{D}$ a estruturas de matrizes de variância-covariância com menos parâmetros.

Dada a simetria da matriz $\mathbf{D}$, é fácil verificar que o número de elementos diferentes da matriz é igual a $\frac{q\times(q+1)}{2}$: q variâncias (elementos diagonais da matriz) e $\frac{q\times(q-1)}{2}$ covariâncias. Estes elementos são os parâmetros de covariância constituintes do vetor $\boldsymbol{\theta}_\mathbf{D}$.

Por exemplo, para um MLM com três efeitos aleatórios associados ao i-ésimo indivíduo, a matriz dos efeitos aleatórios (não estruturada) é:

$$\mathbf{D} = \text{Var}(\mathbf{b}_i) = \begin{bmatrix} \text{Var}(b_{i1}) & \text{Cov}(b_{i1}, b_{i2}) & \text{Cov}(b_{i1}, b_{i3}) \\ \text{Cov}(b_{i1}, b_{i2}) & \text{Var}(b_{i2}) & \text{Cov}(b_{i2}, b_{i3}) \\ \text{Cov}(b_{i1}, b_{i3}) & \text{Cov}(b_{i2}, b_{i3}) & \text{Var}(b_{i3}) \end{bmatrix}.$$

Neste caso, o vetor $\boldsymbol{\theta}_{\mathbf{D}}$ contém $\frac{3\times(3+1)}{2} = 6$ parâmetros de covariância:

$$\boldsymbol{\theta}_{\mathbf{D}} = \begin{bmatrix} \text{Var}(b_{i1}) \\ \text{Cov}(b_{i1}, b_{i2}) \\ \text{Cov}(b_{i1}, b_{i3}) \\ \text{Var}(b_{i2}) \\ \text{Cov}(b_{i2}, b_{i3}) \\ \text{Var}(b_{i3}) \end{bmatrix}.$$

A maior desvantagem de assumir uma matriz de variância-covariância não estruturada é que o número de parâmetros de covariância (comprimento do vetor $\boldsymbol{\theta}_{\mathbf{D}}$) a estimar cresce rapidamente com o número de efeitos aleatórios associados a cada indivíduo. Por exemplo, se considerarmos um MLM com cinco efeitos aleatórios, o número de parâmetros de covariância é 15, enquanto que se considerarmos $q = 10$, cresce para 55. Quando o comprimento do vetor $\boldsymbol{\theta}_{\mathbf{D}}$ é grande em comparação com o tamanho da amostra, a estimação dos parâmetros é instável. Por este motivo, o uso da covariância não estruturada é apelativo apenas nos casos em que o número de indivíduos, N, é bastante maior relativamente ao número de parâmetros de covariância.

Podemos impor outras restrições à matriz $\mathbf{D}$, definindo assim diferentes estruturas. Uma estrutura muito comum é a diagonal, onde os efeitos aleatórios são independentes (elementos de covariância iguais a 0). Regra geral, para a estrutura diagonal, o vetor $\boldsymbol{\theta}_{\mathbf{D}}$ requer q parâmetros de covariância que definem as variâncias (elementos diagonais de $\mathbf{D}$). Num MLM com dois efeitos aleatórios associados ao indivíduo i, a matriz $\mathbf{D}$ com estrutura diagonal tem a seguinte forma:

$$\mathbf{D} = \text{Var}(\mathbf{b}_i) = \begin{bmatrix} \text{Var}(b_{i1}) & 0 \\ 0 & \text{Var}(b_{i2}) \end{bmatrix},$$

e o vetor $\boldsymbol{\theta}_{\mathbf{D}}$ contém dois parâmetros:

$$\boldsymbol{\theta}_{\mathbf{D}} = \begin{bmatrix} \text{Var}(b_{i1}) \\ \text{Var}(b_{i2}) \end{bmatrix}.$$

A matriz $\mathbf{D}$ não estruturada e com estrutura diagonal são as estruturas mais frequentemente usadas. No entanto, existem outras estruturas disponíveis em diferentes softwares.

2.2.2 Método da Máxima Verosimilhança

Tal como o próprio nome indica, o método da Máxima Verosimilhança (MV) permite obter estimadores de parâmetros desconhecidos através da maximização da função de verosimilhança.

No contexto do MLM, construímos a função de verosimilhança a partir da distribuição marginal de $\mathbf{Y}_i|\mathbf{X}_i$, cuja f.d.p. é dada por (2.2). É de notar que os elementos da matriz $\mathbf{V}_i$ são

funções dos parâmetros de covariância contidos em $\boldsymbol{\theta}$.

Assumindo independência entre indivíduos, a função de verosimilhança é:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^{N} f(\mathbf{Y}_i|\mathbf{X}_i) = \prod_{i=1}^{N} (2\pi)^{-\frac{n}{2}} |\mathbf{V}_i|^{-\frac{1}{2}} \exp\left(-\frac{(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})}{2}\right). \tag{2.3}$$

Aplicando o logaritmo a (2.3) obtemos:

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \ln L(\boldsymbol{\beta}, \boldsymbol{\theta}) \\ &= \ln\left(\prod_{i=1}^{N} (2\pi)^{-\frac{n}{2}} |\mathbf{V}_i|^{-\frac{1}{2}} \exp\left(-\frac{(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})}{2}\right)\right) \\ &= \sum_{i=1}^{N} \ln\left((2\pi)^{-\frac{n}{2}} |\mathbf{V}_i|^{-\frac{1}{2}} \exp\left(-\frac{(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})}{2}\right)\right) \\ &= -\frac{N \times n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \ln\left(|\mathbf{V}_i|\right) - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}). \end{aligned}$$

Podemos calcular os estimadores MV para os parâmetros do MLM das seguintes formas:

- **Caso especial: supõe-se $\boldsymbol{\theta}$ conhecido**

Um resultado imediato de assumir $\boldsymbol{\theta}$ conhecido é que a matriz $\mathbf{V}_i$ também passa a ser conhecida. No entanto, esta não é uma situação que ocorra em muitos casos práticos.

Uma vez que $\boldsymbol{\theta}$ é conhecido, os únicos parâmetros a estimar são os coeficientes de regressão $\boldsymbol{\beta}$ (efeitos fixos). Podemos, então, reescrever $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ em função de $\boldsymbol{\beta}$:

$$-\frac{1}{2} \sum_{i=1}^{N} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}). \tag{2.4}$$

Devido ao sinal negativo do termo dependente de $\boldsymbol{\beta}$, maximizar (2.4) em ordem a $\boldsymbol{\beta}$ é equivalente a minimizar:

$$\frac{1}{2} \sum_{i=1}^{N} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}). \tag{2.5}$$

Igualando a zero a derivada de (2.5) em ordem a $\boldsymbol{\beta}$ e resolvendo a equação resultante em ordem ao parâmetro referido, obtemos o estimador MV para $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_{MV} = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{Y}_i\right). \tag{2.6}$$

Demonstração.

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\frac{1}{2} \sum_{i=1}^{N} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})\right) = 0$$

$$
\begin{aligned}
&\iff \frac{\partial}{\partial \boldsymbol{\beta}} \left(\sum_{i=1}^{N} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) \right) = 0 \\
&\iff \frac{\partial}{\partial \boldsymbol{\beta}} \left(\sum_{i=1}^{N} \mathbf{Y}_i^\top \mathbf{V}_i^{-1} \mathbf{Y}_i - \sum_{i=1}^{N} \mathbf{Y}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i\boldsymbol{\beta} - \sum_{i=1}^{N} \boldsymbol{\beta}^\top \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{Y}_i + \sum_{i=1}^{N} \boldsymbol{\beta}^\top \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i\boldsymbol{\beta} \right) = 0 \\
&\iff \frac{\partial}{\partial \boldsymbol{\beta}} \left(- \sum_{i=1}^{N} \mathbf{Y}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i\boldsymbol{\beta} - \sum_{i=1}^{N} \boldsymbol{\beta}^\top \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{Y}_i + \sum_{i=1}^{N} \boldsymbol{\beta}^\top \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i\boldsymbol{\beta} \right) = 0 \\
&\iff - \sum_{i=1}^{N} \mathbf{Y}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i - \sum_{i=1}^{N} \left(\mathbf{X}^\top \mathbf{V}_i^{-1} \mathbf{Y}_i \right)^\top + \sum_{i=1}^{N} \boldsymbol{\beta}^\top \left[\mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i + \left(\mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i \right)^\top \right] = 0 \\
&\iff - \sum_{i=1}^{N} \mathbf{Y}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i - \sum_{i=1}^{N} \mathbf{Y}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i + \boldsymbol{\beta}^\top \sum_{i=1}^{N} \left(\mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i + \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i \right) = 0 \\
&\iff -2 \sum_{i=1}^{N} \mathbf{Y}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i + 2\boldsymbol{\beta}^\top \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i = 0 \\
&\iff \boldsymbol{\beta}^\top \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i = \sum_{i=1}^{N} \mathbf{Y}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i \\
&\iff \left(\boldsymbol{\beta}^\top \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i \right)^\top = \left(\sum_{i=1}^{N} \mathbf{Y}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i \right)^\top \\
&\iff \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i \right) \boldsymbol{\beta} = \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{Y}_i
\end{aligned}
$$

Na situação em que a matriz $\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i$ é invertível, tem-se:

$$
\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{Y}_i \right).
$$

É de notar que, uma vez que só existe solução no caso em que $\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i$ é invertível, não podemos ter multicolinearidade elevada. □

A fórmula descrita em (2.6) define uma relação entre os parâmetros de covariância, $\boldsymbol{\theta}$, e o valor de $\boldsymbol{\beta}$ que maximiza a função $l(\boldsymbol{\beta}, \boldsymbol{\theta})$. De seguida, iremos usar esta relação para estimar os efeitos fixos quando $\boldsymbol{\theta}$ é desconhecido.

- **Caso geral: $\boldsymbol{\theta}$ é desconhecido**

Na maioria dos casos práticos $\boldsymbol{\theta}$ é desconhecido. Nestas situações, o estimador MV de $\boldsymbol{\theta}$ é obtido maximizando a função $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ com respeito a $\boldsymbol{\theta}$, após $\boldsymbol{\beta}$ ter sido substituído por (2.6). Em geral, maximizar $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ com respeito a $\boldsymbol{\theta}$ requer processos numéricos de otimização não linear, com restrições de desigualdade impostas a $\boldsymbol{\theta}$, garantindo assim que as matrizes $\mathbf{D}$ e $\boldsymbol{\Sigma}_i$ são definidas positivas. Não existe, no entanto, uma solução ótima para $\boldsymbol{\theta}$, pelo que a sua estimativa é obtida realizando iterações computacionais até se obter convergência.

Após a estimativa MV dos parâmetros de covariância (e, consequentemente, estimativas dos elementos das matrizes $\mathbf{D}$ e $\boldsymbol{\Sigma}_i$), podemos calcular $\widehat{\mathbf{V}}_{iMV}$ substituindo as matrizes $\mathbf{D}$ e $\boldsymbol{\Sigma}_i$ pelos seus estimadores MV, $\widehat{\mathbf{D}}_{MV}$ e $\widehat{\boldsymbol{\Sigma}}_{iMV}$:

$$\widehat{\mathbf{V}_{iMV}} = \mathbf{Z}_i \widehat{\mathbf{D}}_{MV} \mathbf{Z}_i^\top + \widehat{\mathbf{\Sigma}_{iMV}}. \tag{2.7}$$

De seguida, substituimos $\mathbf{V}_i$ por (2.7) na equação (2.6), de forma a obter o estimador MV de $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_{MV} = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \widehat{\mathbf{V}}_{iMV}^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \widehat{\mathbf{V}}_{iMV}^{-1} \mathbf{Y}_i \right). \tag{2.8}$$

No método MV os estimadores dos parâmetros de covariância são enviesados devido à perda de graus de liberdade resultante da estimação dos efeitos fixos (Verbeke & Molenberghs (2000)). Este problema é resolvido fazendo uso do método da Máxima Verosimilhança Restrita (MVR) para estimar as componentes de covariância.

2.2.3 Método da Máxima Verosimilhança Restrita

O método MVR é uma forma alternativa de obter estimadores centrados para os parâmetros de covariância contidos em $\boldsymbol{\theta}$. Este método foi introduzido por Patterson & Thompson (1971) e desenvolvido posteriormente por Harville (1977).

A ideia principal por trás do método MVR é separar a parte dos dados usada para a estimativa de $\mathbf{V}_i$ daquela usada para estimar $\boldsymbol{\beta}$, sendo que $\mathbf{V}_i$ apenas é estimada com base na parte relevante dos dados. Ou seja, a ideia fundamental é eliminar $\boldsymbol{\beta}$ da função de verosimilhança logarítmica ($l(\boldsymbol{\beta}, \boldsymbol{\theta})$), de forma a esta ser escrita, exclusivamente, em termos de $\mathbf{V}_i$. Isto é conseguido maximizando a verosimilhança de uma transformação linear dos dados, $\mathbf{U} = \mathbf{A}^\top \mathbf{Y}$, onde $\mathbf{A}$ é uma matriz com dimensão $(N \times n) \times (N \times n - p)$, ortogonal às colunas de $\mathbf{X}$. Em geral, obtém-se a verosimilhança restrita utilizando a matriz de projeção ortogonal que gera os resíduos do ajuste obtido pelo método dos mínimos quadrados, $\mathbf{A} = \mathbf{I} - \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top$. A verosimilhança resultante desta transformação não depende dos efeitos fixos (uma vez que a distribuição de $\mathbf{U}$ não depende de $\boldsymbol{\beta}$), nem da matriz $\mathbf{A}$ escolhida, sendo que se obtêm sempre os mesmos estimadores dos parâmetros de covariância.

Em vez de maximizar $l(\boldsymbol{\beta}, \boldsymbol{\theta})$, o método MVR para a estimativa de $\boldsymbol{\theta}$ é baseado na otimização da seguinte função de verosimilhança logarítmica:

$$\begin{aligned} l_{MVR}(\boldsymbol{\theta}) = & -\frac{1}{2}(N \times n - p) \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \ln(|\mathbf{V}_i|) - \frac{1}{2} \sum_{i=1}^{N} \left(\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta} \right)^\top \mathbf{V}_i^{-1} \left(\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta} \right) - \\ & - \frac{1}{2} \sum_{i=1}^{N} \ln \left(\left| \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i \right| \right). \end{aligned}$$

Ao contrário do método MV, o método MVR não fornece uma estimativa para os coeficientes de regressão contidos em $\boldsymbol{\beta}$: uma vez estimada a matriz $\mathbf{V}_i$ (maximizando $l_{MVR}(\boldsymbol{\theta})$ com respeito a $\boldsymbol{\theta}$), determinamos o estimador MVR dos efeitos fixos através da equação (2.8) definida para o método MV.

Apesar de usarmos a fórmula obtida para o estimador MV de $\boldsymbol{\beta}$ no cálculo de $\widehat{\boldsymbol{\beta}}_{MVR}$, é importante notar que os estimadores resultantes são diferentes, pois a matriz $\widehat{\mathbf{V}_i}$ é diferente em

ambos os métodos.

2.2.4 Máxima Verosimilhança vs. Máxima Verosimilhança Restrita

Existem três grandes diferenças entre os dois métodos apresentados:

- O método MV não tem em consideração a perda de graus de liberdade resultante da estimação de $\boldsymbol{\beta}$, gerando assim estimadores enviezados para os parâmetros de covariâcia. Este problema é solucionado pelo método MVR;
- O método MV produz estimadores para os efeitos fixos, enquanto que o método MVR fornece primeiro estimadores para os parâmetros de covariância e usa-os posteriormente para calcular os estimadores para $\boldsymbol{\beta}$ através da fórmula resultante do método MV;
- Uma diferença importante entre a função log-verosimilhança e a função log-verosimilhança restrita é o termo $\frac{1}{2}\sum_{i=1}^{N}\ln\left(\left|\mathbf{X}_i^\top\mathbf{V}_i^{-1}\mathbf{X}_i\right|\right)$, do qual podemos concluir que o método MVR produz estimadores que variam face a reparametrizações dos efeitos fixos. Como consequência, modelos lineares mistos com diferentes estruturas de efeitos fixos não são comparáveis com base na função de verosimilhança restrita. Em particular, testes de razão de verosimilhança não são válidos sempre que a estimação tiver sido feita pelo método MVR.

Apesar destas diferenças, ambos os métodos (MV e MVR) produzem, na maioria dos casos, estimativas semelhantes, sendo que a diferença cresce com o aumento do número de termos fixos do modelo (Zuur *et al.*, 2009).

Os dois métodos envolvem algoritmos de otimização complexos, que normalmente requerem valores iniciais para os parâmetros a estimar e várias iterações subsequentes para encontrarem os valores dos parâmetros que maximizam a função de verosimilhança. Na próxima secção iremos apresentar alguns destes algoritmos.

2.3 Algoritmos de Otimização

A principal dificuldade na análise do MLM é estimar os parâmetros de covariância. Este processo computacional requer métodos numéricos de otimização da função log-verosimilhança (introduzida na secção 2.2.2 para o método MV, e na secção 2.2.3 para o método MVR), em virtude das restrições não lineares impostas ao parâmetro $\boldsymbol{\theta}$ de forma a assegurar que as matrizes $\mathbf{D}$ e $\boldsymbol{\Sigma}_i$ são definidas positivas.

Os métodos numéricos de otimização mais utilizados, e que se encontram implementados nos procedimentos de cálculo para o MLM em programas estatísticos, são o algoritmo de estimação-maximização[1] (EM) e o método de Newton-Raphson (NR).

Ambos os métodos apresentados requerem valores iniciais para os parâmetros a estimar, que podem ser obtidos com base em estudos anteriores ou a partir dos próprios dados, e um valor, $\epsilon > 0$, para definir o critério de convergência.

[1] Expectation-Maximization

2.3.1 Algoritmo EM

O algoritmo EM (Dempster *et al.*, 1977) é um processo iterativo popular para modelos com dados incompletos. Cada iteração do modelo envolve dois passos:

- **Passo E**: cálculo de $\mathrm{E}\left[l(\boldsymbol{\theta})_{\mathbf{b}|\mathbf{Y}}\right]$

Usamos os parâmetros de covariância de cada iteração para avaliar a distribuição de $\mathbf{b}|\mathbf{Y}$ e portanto determinar o valor esperado da função de versomilhança para um novo valor de $\boldsymbol{\theta}$, dada a distribuição condicional.

Uma vez que utiliza a informação disponível a cada iteração (valores atuais das estimativas dos parâmetros de covariância e valores observados da variável dependente), podemos dizer que o algoritmo EM simula a estimativa que seria possível obter se tivéssemos dados completos.

- **Passo M**: maximização de $\mathrm{E}\left[l(\boldsymbol{\theta})_{\mathbf{b}|\mathbf{Y}}\right]$

Consiste em maximizar a função log-verosimilhança (calculada no passo anterior) com respeito a $\boldsymbol{\theta}$, de forma a obter o estimador do parâmetro de covariância para a iteração seguinte.

O pressuposto subjacente deste algoritmo é que otimizar a função log-verosimilhança para os dados "completos" é mais simples que para os dados observados. Pode-se mostrar que cada iteração do algoritmo resulta num aumento da função log-verosimilhança (Dempster *et al.*, 1977).

A principal desvantagem do algoritmo EM é que pode não atingir convergência no caso dos valores iniciais serem arbitrários. Para além disso, a precisão dos estimadores é excessivamente otimista, pois a estimativa é baseada na verosimilhança do último passo M, que usa os dados "completos" em vez dos observados.

2.3.2 Algoritmo Newton-Raphson

O algoritmo NR (Thisted, 1988), bem como outros que são variações do original, são dos métodos de otimização mais utilizados para o MLM.

Este algoritmo minimiza uma função objetivo que é o produto de -2 pela função log-verosimilhança (especificadas na secção 2.2.2 para o método MV e na secção 2.2.3 para o método MVR). A cada iteração, o algoritmo NR requer o cálculo do vetor das derivadas parciais da função log-verosimilhança com respeito aos parâmetros de covariância (gradiente), assim como a matriz da segunda derivada (matriz Hessiana). Ao igualar o gradiente a $\mathbf{0}$ obtemos um sistema de equações (geralmente não lineares), que é necessário resolver para encontrar os estimadores dos parâmetros contidos em $\boldsymbol{\theta}$.

O cálculo da matriz Hessiana a cada iteração requer algum tempo de cálculo, no entanto, a convergência é, na maioria dos casos, atingida mais rapidamente que no algoritmo EM.

2.3.3 EM vs. NR

As iterações individuais do algoritmo EM são fáceis e rápidas de calcular. No entanto, apesar das iterações iniciais se aproximarem rapidamente do ótimo, é de lenta convergência quando se aproxima deste. Sendo que, por vezes, é difícil decidir se a iteração terminou ou não. Por outro

lado, as iterações de Newton-Raphson são mais complexas, podendo ser muito instáveis quando longe do ótimo. Contudo, quando próximo deste, convergem rapidamente.

Por estes motivos, é útil usar uma abordagem híbrida, realizando um número moderado de iterações do algoritmo EM com um $\boldsymbol{\theta}_0$ (inicial) e, quando perto do ótimo, passar para o método NR. A função *lme*, da biblioteca *nlme* do R (usada para este estudo), implementa este esquema de otimização híbrida, onde, por defeito, realiza 25 iterações do algoritmo EM antes de mudar para o método NR.

2.4 Predição dos efeitos aleatórios: Melhor Preditor Linear Centrado (BLUP[2])

Os valores dos vários efeitos aleatórios, $\mathbf{b}_1, \ldots, \mathbf{b}_q$, não são parâmetros desconhecidos e fixos que possam ser estimados, como é o caso de $\boldsymbol{\beta}$. Ao invés, são variáveis aleatórias, que refletem o desvio na evolução de cada indivíduo em relação ao valor esperado da população ($\mathbf{X}_i\boldsymbol{\beta}$). Como resultado, falamos em prever os efeitos aleatórios em vez de estimar.

Uma vez que assumimos em (2.1) que o valor esperado da distribuição normal multivariada de $\mathbf{b}_i$ é $\mathbf{0}$, não é do nosso interesse estimar o valor esperado do conjunto dos efeitos aleatórios para o i-ésimo indivíduo. Contudo, em geral, prever uma variável aleatória é equivalente a prever o seu valor esperado condicional aos dados disponíveis. Assim sendo, o melhor preditor de $\mathbf{b}_i$ é $\mathrm{E}(\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i)$. Ora, dado que

$$\begin{aligned} &\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \\ &\iff \begin{bmatrix} \mathbf{Y}_i \\ \mathbf{b}_i \end{bmatrix} = \begin{bmatrix} \mathbf{X}_i\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_i & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}_i \\ \mathbf{e}_i \end{bmatrix}, \end{aligned}$$

podemos escrever

$$\begin{bmatrix} \mathbf{Y}_i \\ \mathbf{b}_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mathrm{E}(\mathbf{Y}_i) \\ \mathrm{E}(\mathbf{b}_i) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_i\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathrm{Var}(\mathbf{Y}_i) & \mathrm{Cov}(\mathbf{Y}_i, \mathbf{b}_i) \\ \mathrm{Cov}(\mathbf{b}_i, \mathbf{Y}_i) & \mathrm{Var}(\mathbf{b}_i) \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \boldsymbol{\Sigma}_i & \mathbf{Z}_i\mathbf{D} \\ \mathbf{D}\mathbf{Z}_i^\top & \mathbf{D} \end{bmatrix} \right),$$

de onde se conclui

$$\mathbf{b}_i|\mathbf{y}_i \sim N\left(\mathbf{0} + \mathbf{D}\mathbf{Z}_i^\top (\underbrace{\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \boldsymbol{\Sigma}_i}_{=\mathbf{V}_i})^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \mathbf{D} - \mathbf{D}\mathbf{Z}_i^\top(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \boldsymbol{\Sigma}_i)^{-1}\mathbf{Z}_i\mathbf{D} \right).$$

Então, o melhor preditor linear centrado (BLUP) de $\mathbf{b}_i$ é dado por:

$$\widehat{\mathbf{b}_i} = \mathrm{E}(\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i) = \mathbf{D}\mathbf{Z}_i^\top \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}). \tag{2.9}$$

Este preditor dos efeitos aleatórios depende dos parâmetros de covariância. Quando $\boldsymbol{\theta}$ é

[2]Best Linear Unbiased Estimator.

conhecido, podemos substituir em (2.9) os estimadores (MV ou MVR) dos parâmetros de covariância, obtendo assim o denominado BLUP empírico (EBLUP):

$$\widehat{\mathbf{b}_i} = \widehat{\mathbf{D}}\mathbf{Z}_i^\top \widehat{\mathbf{V}_i}^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}).$$

Os melhores preditores lineares centrados são:

- Lineares, porque são funções lineares dos dados observados, $\mathbf{y}_i$;
- Centrados, pois o seu valor esperado é igual ao valor esperado dos efeitos aleatórios para o i-ésimo indivíduo ($\mathrm{E}(\widehat{\mathbf{b}_i} - \mathbf{b}_i) = 0$);
- "Melhores" (mais eficientes), na medida em que têm a menor variância entre todos os preditores lineares centrados ($\mathrm{Var}(\widehat{\mathbf{b}_i} - \mathbf{b}_i)$ não é maior que $\mathrm{Var}(\mathbf{u}_i - \mathbf{b}_i)$, onde $\mathbf{u}_i$ é um preditor linear centrado qualquer).

2.5 Inferência Estatística

Depois de ajustarmos um modelo estatístico aos dados, é necessário avaliar a precisão das estimativas, a significância dos vários termos no modelo, assim como comparar o ajustamento entre modelos. Para tal, nesta secção apresentamos testes de hipótese para os parâmetros do MLM.

2.5.1 Testes da Razão de Verosimilhanças para os Efeitos Fixos

Um método geral para comparação de modelos aninhados, ajustados pelo método da máxima verosimilhança, é o teste da razão de verosimilhanças (Lehnam, 1986). Este teste também pode ser usado quando os modelos são ajustados pelo método MVR, mas apenas no caso de ambos os modelos terem a mesma estrutura de efeitos fixos.

Antes de avançarmos, é importante definir o conceito de modelo aninhado, pelo que vamos assumir que queremos comparar dois modelos: M_1 e M_2. Dizemos que M_1 é um modelo aninhado de M_2, se o espaço de parâmetros do modelo M_1 é um subespaço do espaço de parâmetros de M_2. Podemos dizer, com menos formalidade, que os parâmetros do modelo aninhado podem ser obtidos impondo certas restrições aos parâmetros do modelo mais geral.

Seja, portanto, L_2 a verosimilhança do modelo geral (M_2) e L_1 a verosimilhança do modelo aninhado (M_1). Então, temos de ter $L_2 > L_1$ e, consequentemente, $\ln(L_2) > \ln(L_1)$. A estatística de teste é dada por:

$$2\ln\left(\frac{L_2}{L_1}\right) = 2\left[\ln(L_2) - \ln(L_1)\right]. \tag{2.10}$$

Se denotarmos por k_i o número de parâmetros a estimar do modelo i, então, a distribuição assintótica da estatística de teste, sob a hipótese nula do modelo mais restrito ser adequado, é a distribuição χ^2 com $k_2 - k_1$ graus de liberdade.

Usando o resultado da equação (2.10), podemos testar hipóteses para os parâmetros do MLM. Se a estatística de teste for suficientemente grande, então há evidências contra a hipótese nula

e a favor do modelo de referência (modelo geral). Se as verosimilhanças tomarem valores muito próximos, resultando numa estatística de teste pequena, então temos evidências a favor do modelo aninhado (hipótese nula). Uma desvantagem de realizar testes de razão de verosimilhanças nestas circunstâncias é que tendem a ser anti-conservativos. Isto é, o valor-p resultante da distribuição $\chi^2_{k_2-k_1}$ é, geralmente, inferior ao verdadeiro valor-p do teste. Segundo Stram & Lee (1994), esta tendência é consequência das restrições impostas ao modelo geral, que envolvem definir como zero a variância de certas componentes dos efeitos aleatórios, estando assim no limite da região dos parâmetros. Por outro lado, à medida que se aumenta o número de parâmetros removidos dos efeitos fixos, em comparação com o número total de observações, a imprecisão dos valores-p reportados pode ser substancial (Pinheiro & Bates, 2000). Outros testes alternativos para avaliar a significância dos efeitos fixos podem ser utilizados, sendo que os mais recomendados são os testes-t e os testes-F aproximados.

Testes Alternativos para os Efeitos Fixos

O **teste-t** aproximado é muitas vezes usado para avaliar a significância marginal de cada efeito fixo singular, quando todos os restantes coeficientes dos efeitos fixos estão presentes no modelo:

$$H_0 : \beta_i = 0 \quad vs. \quad H_1 : \beta_i \neq 0, \quad i = 1, \ldots, p.$$

A estatística do teste-t correspondente é calculada da seguinte forma:

$$t = \frac{\widehat{\beta}_i}{se\left(\widehat{\beta}_i\right)},$$

que, sob a hipótese nula, tem distribuição assintótica t de Student com gl_j graus de liberdade, e onde $se(.)$ representa o erro padrão (desvio padrão da distribuição amostral da estatística de teste) e é calculado através da variância do estimador $\widehat{\boldsymbol{\beta}}$ dado em (2.8). Equivalentemente, a estatística t^2 segue a distribuição $F(1, gl_j)$. O teste-t é aproximado uma vez que a estatística de teste está condicionada ao estimador MVR do vetor dos parâmetros de covariância.

Os **testes-F** são usados para testar hipósteses lineares sobre múltiplos efeitos fixos. Por exemplo, podemos querer testar se algum dos parâmetros associados aos níveis de um fator fixo é diferente de zero. Em geral, ao testar uma hipótese linear da forma

$$H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0} \quad vs. \quad H_1 : \mathbf{L}\boldsymbol{\beta} \neq \mathbf{0},$$

em que $\mathbf{L}$ é uma matriz conhecida, a estatística-F é definida por:

$$F = \frac{\widehat{\boldsymbol{\beta}}^\top \mathbf{L}^\top \left(\mathbf{L} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \widehat{\mathbf{V}_i}^{-1} \mathbf{X}_i \right)^{-1} \mathbf{L}^\top \right)^{-1} \mathbf{L}\widehat{\boldsymbol{\beta}}}{\text{rank}(\mathbf{L})},$$

que, sob a hipótese nula, tem distribuição assintótica F com $(\text{rank}(\mathbf{L})^3, v)$ graus de liberdade.

[3] $\text{rank}(\mathbf{L})$ representa a característica da matriz $\mathbf{L}$.

Para ambos os testes o número de graus de liberdade do denominador (gl_i para o teste-t e v para o teste-F) têm de ser estimados a partir dos dados. Existem diversos métodos para estimar o número de graus de liberdade, sendo que diferentes métodos conduzem a resultados diferentes. Contudo, uma vez que estamos a trabalhar com dados longitudinais, onde cada indivíduo contribui com informação independente, o número de graus de liberdade é suficientemente grande (qualquer que seja o método usado para a estimativa), o que resulta em valores-p muito semelhantes (Verbeke & Molenberghs, 2000).

2.5.2 Testes da Razão de Verosimilhanças para os Parâmetros de Covariância

Normalmente, os testes da razão de verosimilhanças são usados para testar a significância dos termos na estrutura dos efeitos aleatórios. Isto é, ajustamos modelos aninhados com efeitos aleatórios diferentes e aplicamos estes testes para os comparar. Contudo, temos de ter em consideração que as restrições impostas aos parâmetros de covariância podem resultar numa má especificação do modelo, conduzindo a inferências erradas sobre os dados a estudar. Assim sendo, é do nosso interesse saber quais os efeitos aleatórios a incluir no modelo, bem como a estrutura da sua matriz de variância-covariância, $\mathbf{D}$.

Para a construção de testes da razão de verosimilhanças para os parâmetros de covariância devemos usar os seus respetivos estimadores MVR, pois, como já foi referido anteriormente, o método MV não produz estimadores centrados para estes parâmetros. Assumimos que ambos os modelos a comparar (modelo aninhado e de referência) têm a mesma estrutura de efeitos fixos, mas diferentes estruturas de covariância.

Para um dado $\boldsymbol{\beta}$, podemos testar várias hipóteses, como por exemplo, testar a hipótese nula de haver q efeitos aleatórios *vs.* a hipótese alternativa de haver $q + k$, ou testar a hipótese nula de os efeitos aleatórios serem independentes *vs.* a hipótese alternativa de não serem. A distribuição nula da estatística de teste vai depender da possibilidade dos valores dos parâmetros de covariância da hipótese nula pertencerem à fronteira do espaço de parâmetros.

- **Caso 1: Os parâmetros de covariância que satisfazem a hipótese nula não pertencem à fronteira do espaço de parâmetros**

Seja k_1 o número de parâmetros de covariância do modelo aninhado, e k_2 o número de parâmetros de covariância do modelo de referência.

No caso em que a hipótese nula não envolve parâmetros contidos na fronteira do espaço de parâmetros (como por exemplo, testar um modelo com variância residual heterogénea *vs.* um modelo com variância residual constante, ou testar se a covariância entre dois efeitos aleatórios é igual a zero), a estatística de teste tem distribuição assintótica $\chi^2_{k_2-k_1}$.

- **Caso 2: Os parâmetros de covariância que satisfazem a hipótese nula pertencem à fronteira do espaço de parâmetros**

Testes de hipóteses em que os valores dos parâmetros de covariância da hipótese nula pertencem à fronteira do espaço de parâmetros surgem no contexto de avaliar se um efeito aleatório deve ou não ser incluído no modelo. Não testamos diretamente hipóteses sobre os efeitos aleatórios em si, mas sim a possibilidade de as respetivas variâncias e covariâncias serem igual a zero.

No caso em queremos testar a hipótese nula de incluir q efeitos aleatórios *vs.* a hipótese alternativa de incluir $q+1$ efeitos aleatórios, a estatística de teste tem distribuição assintótica $0.5\chi^2_q + 0.5\chi^2_{q+1}$ (Verbeke e Molenberghs, 2000). Contudo, segundo Pinheiro & Bates (2000), este ajustamento não é sempre bem sucedido, pelo que recomendam a abordagem *naive* implementada na biblioteca *nlme* do R, que usa a distribuição χ^2 com o número de graus de liberdade determinado pela diferença entre o número de parâmetros estimados nos modelos das hipóteses nula e alternativa. É de notar que o valor-p resultante é conservativo, isto é, em alguns casos pode ser maior que o verdadeiro *p-value* (Pinheiro & Bates, 2000).

2.5.3 Critérios de Informação

Outro conjunto de ferramentas útil na seleção de modelos são os critérios de informação, que fornecem alternativas para avaliar o ajustamento de modelos com base na maximização da função log-verosimilhança, penalizando aqueles com mais parâmetros. Uma característica chave dos critérios de informação abordados nesta secção é que estes permitem comparar quaisquer dois modelos ajustados ao mesmo conjunto de dados, isto é, não é necessário que ambos os modelos tenham a mesma estrutura de efeitos fixos. Quanto menor for o valor do critério de informação, melhor é o ajustamento.

O **Critério de Informação de Akaike** (AIC), proposto por Akaike (1974), pode ser calculado com base na função log-verosimilhança (MV ou MVR), através da seguinte expressão:

$$\text{AIC} = -2 \times l\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}\right) + 2n_{par},$$

onde n_{par} representa o número total de parâmetros a estimar no modelo para ambos os efeitos fixos e aleatórios.

O **Critério de Informação Bayesiana** (BIC), proposto por Schwarz (1978), também é um critério de informação usual, que pode ser calculado da seguinte forma:

$$\text{BIC} = -2 \times l\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}\right) + n_{par} \ln(N \times n),$$

sendo n_{par} definido como anteriormente, e $N \times n$ o número total de observações usados no modelo ajustado.

Ambos os critérios são semelhantes, sendo que o BIC é mais sensível ao número de parâmetros incluídos no modelo, penalizando ainda mais, em comparação com o AIC, o modelo com mais parâmetros.

É de notar que a comparação de modelos ajustados pelo método MVR só pode ser feita na condição de ambos os modelos terem a mesma estrutura de efeitos fixos.

Trabalhos recentes (Gurka, 2006) sugerem que nenhum dos critérios se destaca como sendo o melhor a usar, e que ainda há muito para compreender sobre o papel que os critérios de informação têm na seleção de modelos.

2.6 Diagnóstico

Após ajustar um MLM aos dados, e antes de fazer inferências sobre o modelo ajustado, é importante realizar um diagnóstico para verificar se as premissas subjacentes são válidas.

Os métodos de diagnóstico para os modelos lineares usuais estão bem estabelecidos na literatura estatística. Contudo, o diagnóstico de modelos lineares mistos é mais difícil de realizar e interpretar devido à complexidade do modelo e à presença dos efeitos aleatórios e das várias estruturas de covariância consideradas.

Num MLM existem dois pressupostos que têm de ser verificados:

- **Premissa 1**: Os erros aleatórios, $\mathbf{e}_i$, são independentes (para o mesmo indivíduo) e identicamente distribuídos. Seguem uma distribuição normal com valor médio nulo e variância constante σ^2. São, também, independentes dos efeitos aleatórios;
- **Premissa 2**: Os efeitos aleatórios, $\mathbf{b}_i$, seguem uma distribuição normal multivariada, com média $\mathbf{0}$ e matriz de variância-covariância $\mathbf{D}$.

A biblioteca *nlme* do R fornece vários métodos para verificar a validade destes pressupostos. Os métodos mais úteis são baseados em gráficos dos resíduos, dos valores ajustados e dos efeitos aleatórios. A validade dos pressupostos relativos às distribuições, quer dos efeitos aleatórios, quer dos erros aleatórios, também podem ser formalmente avaliadas através de testes de hipóteses.

Em geral, o diagnóstico do modelo deve fazer parte do processo de construção. Nesta secção vamos considerar o diagnóstico apenas para o modelo ajustado final por uma questão de simplicidade.

2.6.1 Verificação das Condições sobre os Erros Aleatórios

Normalmente, o diagnóstico dos erros aleatórios (para o mesmo indivíduo) envolve a avaliação dos pressupostos de normalidade, de média zero e de variância constante. A dependência entre erros aleatórios para o mesmo indivíduo é, em geral, modelada através de estruturas de correlação, que são discutidas na secção 2.7, onde são descritos métodos para avaliar a premissa de independência.

As quantidades primárias para a verificação da premissa 1 são os resíduos individuais, definidos como a diferença entre a resposta observada e o valor ajustado para cada uma das n observações, para cada indivíduo. Os resíduos individuais, condicionados aos parâmetros de covariância, são as estimativas BLUP dos erros aleatórios, uma vez que os parâmetros de covariância têm de ser substituídos pelas suas estimativas. No entanto, fornecem bons estimadores para $\mathbf{e}_i$ e podem ser usados para verificar a validade da premissa 1 (Pinheiro & Bates, 2000).

Podemos escolher entre os seguintes tipos de resíduos: resíduos correntes (*raw residuals*), resíduos padronizados (de Pearson) e resíduos normalizados. É habitual fazer-se uso dos seguintes gráficos para a validação dos pressupostos impostos aos erros aleatórios:

- Caixas de bigodes dos resíduos individuais;
- Gráfico dos resíduos de Pearson *vs.* valores ajustados;

- Gráfico dos valores observados *vs.* valores ajustados;
- Gráfico da função de autocorrelação empírica;
- Gráfico quantil-quantil (QQ) da distribuição normal *vs.* resíduos.

2.6.2 Verificação das Condições sobre os Efeitos Aleatórios

Nesta secção vamos descrever os principais métodos de diagnóstico para avaliar a premissa 2. A escolha natural para diagnosticar os pressupostos impostos à distribuição dos efeitos aleatórios são as suas estimativas EBLUP.

Os gráficos mais usados para a verificação das hipóteses subjacentes aos efeitos aleatórios são:

- Gráfico QQ da distribuição normal *vs.* estimativas EBLUP;
- Histograma das estimativas EBLUP;
- Gráfico de dispersão das estimativas EBLUP.

Contudo, a menos que as matrizes $\mathbf{X}_i$ e $\mathbf{Z}_i$ sejam as mesmas, os estimadores $\widehat{\mathbf{b}_i}$ têm todos distribuições diferentes, o que pode resultar numa má interpretação dos gráficos relativos ao papel de probabilidades e histogramas das estimativas EBLUP não padronizados. Sugere-se, então, verificar o pressuposto da distribuição Gaussiana através dos EBLUP padronizados (DeGruttola *et al.*, 1991).

Em geral, a distribuição das estimativas EBLUP não reflete necessariamente a verdadeira distribuição dos efeitos aleatórios, sendo que verificar a validade da premissa 2 nestas circunstâncias tem pouco valor (West *et al.*, 2007).

No entanto, segundo Verbeke & Molenberghs (2000), mesmo que não se verifique a validade da hipótese da distribuição de $\mathbf{b}_i$ ser Gaussiana, se o interesse recair apenas na parte fixa do modelo, as inferências resultantes são válidas.

2.7 Modelação da Estrutura dos Erros Aleatórios

O MLM enunciado em (2.1), também denominado Modelo Linear Misto Básico (MLMB), permite uma certa flexibilidade na especificação da estrutura dos efeitos aleatórios, $\mathbf{b}_i$, mas limita muito a estrutura dos erros aleatórios, $\mathbf{e}_i$, ao impor que estes sejam independentes, e variáveis aleatórias identicamente distribuídas com média zero e variância constante, ou seja

$$\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}. \tag{2.11}$$

Apesar do MLMB permitir a modelação de um elevado número de situações práticas, a restrição (2.11) não permite modelar situações em que a variância intra-indivíduos não é constante, isto é, situações em que os erros aleatórios são heterocedastáticos, ou em que os erros aleatórios estão correlacionados.

Neste capítulo vamos extender o MLMB e introduzir o **Modelo Linear Misto Geral** (MLMG), cujas diferentes estruturas da matriz $\mathbf{\Sigma}_i$ (também apresentadas nesta secção) permitem modelar não só a heterocedasticidade, como a correlação dos erros aleatórios.

2.7.1 Modelo Linear Misto Geral

Vamos considerar um modelo mais geral, onde a reta de regressão permanece a mesma, mas com a seguinte generalização da distribuição dos erros aleatórios:

$$\begin{aligned} &\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \\ &\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}), \\ &\mathbf{e}_i \sim N(\mathbf{0}, \sigma^2\mathbf{\Lambda}_i), \end{aligned} \tag{2.12}$$

onde $\mathbf{\Lambda}_i$ é uma matriz $n \times n$, definida positiva, que depende de i somente através da sua dimensão, e é parametrizada por um conjunto, geralmente pequeno, de parâmetros $\boldsymbol{\lambda}$ (Pinheiro & Bates, 2000). Tal como no MLMB, os erros aleatórios, $\mathbf{e}_i$, são independentes para diferentes indivíduos e independentes dos efeitos aleatórios, $\mathbf{b}_i$.

Uma vez que $\mathbf{\Lambda}_i$ é uma matriz definida positiva, segundo Thisted (1988), admite raiz quadrada invertível $\mathbf{\Lambda}_i^{1/2}$, com inversa $\mathbf{\Lambda}_i^{-1/2}$, tal que:

$$\mathbf{\Lambda}_i = \left(\mathbf{\Lambda}_i^{1/2}\right)^\top \mathbf{\Lambda}_i^{1/2},$$

e

$$\mathbf{\Lambda}_i^{-1} = \mathbf{\Lambda}_i^{-1/2}\left(\mathbf{\Lambda}_i^{-1/2}\right)^\top.$$

Considerando a seguinte reparametrização do modelo

$$\begin{aligned} \mathbf{Y}_i^* &= \left(\mathbf{\Lambda}_i^{-1/2}\right)^\top \mathbf{Y}_i, \\ \mathbf{X}_i^* &= \left(\mathbf{\Lambda}_i^{-1/2}\right)^\top \mathbf{X}_i, \\ \mathbf{Z}_i^* &= \left(\mathbf{\Lambda}_i^{-1/2}\right)^\top \mathbf{Z}_i, \\ \mathbf{e}_i^* &= \left(\mathbf{\Lambda}_i^{-1/2}\right)^\top \mathbf{e}_i, \end{aligned}$$

com

$$\begin{aligned} \mathrm{E}(\mathbf{e}_i^*) &= \mathrm{E}\left[\left(\mathbf{\Lambda}_i^{-1/2}\right)^\top \mathbf{e}_i\right] = \left(\mathbf{\Lambda}_i^{-1/2}\right)^\top \underbrace{\mathrm{E}(\mathbf{e}_i)}_{=0} = 0, \\ \mathrm{Var}(\mathbf{e}_i^*) &= \mathrm{Var}\left[\left(\mathbf{\Lambda}_i^{-1/2}\right)^\top \mathbf{e}_i\right] = \mathbf{\Lambda}_i^{-1/2} \underbrace{\mathrm{Var}(\mathbf{e}_i)}_{=\sigma^2\mathbf{\Lambda}_i} \left(\mathbf{\Lambda}_i^{-1/2}\right)^\top = \sigma^2\mathbf{\Lambda}_i^{-1/2} \underbrace{\mathbf{\Lambda}_i \left(\mathbf{\Lambda}_i^{-1/2}\right)^\top}_{=\mathbf{\Lambda}_i^{1/2}} = \sigma^2\mathbf{I}, \end{aligned}$$

podemos afirmar que $\mathbf{e}_i^* \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, e reescrever o modelo (2.12) da seguinte forma:

$$\begin{aligned}
&\mathbf{Y}_i^* = \mathbf{X}_i^*\boldsymbol{\beta} + \mathbf{Z}_i^*\mathbf{b}_i + \mathbf{e}_i^*,\\
&\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}),\\
&\mathbf{e}_i^* \sim N(\mathbf{0}, \sigma^2\mathbf{I}),
\end{aligned}$$

de onde podemos concluir que $\mathbf{Y}_i^*$ é descrito por um MLMB.

É de notar que, para uma amostra $\mathbf{y}$ dada, a derivada da transformação linear $\mathbf{y}_i^* = \left(\boldsymbol{\Lambda}_i^{-1/2}\right)^\top \mathbf{y}_i$ é, simplesmente, $d\mathbf{y}_i^* = \left|\left(\boldsymbol{\Lambda}_i^{-1/2}\right)^\top\right| d\mathbf{y}_i = \left|\boldsymbol{\Lambda}_i^{-1/2}\right| d\mathbf{y}_i$. Logo, a função de verosimilhança L, para o MLMG descrito em (2.12), é dada por:

$$\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}; \mathbf{y}) &= \prod_{i=1}^{N} f(\mathbf{y}_i; \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda})\\
&= \prod_{i=1}^{N} f(\mathbf{y}_i^*; \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \left|\boldsymbol{\Lambda}_i^{-1/2}\right|\\
&= L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}; \mathbf{y}^*) \prod_{i=1}^{N} \left|\boldsymbol{\Lambda}_i^{-1/2}\right|.
\end{aligned}$$

A função de verosimilhança $L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}; \mathbf{y}^*)$ corresponde à função de verosimilhança apresentada na secção 2.2.2 para o MLMB, pelo que todos os resultados obtidos nessa secção são válidos.

Analogamente, os resultados obtidos na secção 2.2.3 são válidos para a função de máxima verosimilhança restrita $L_{MVR}(\boldsymbol{\theta}, \boldsymbol{\lambda}; \mathbf{y}^*)$ e, consequentemente, podem ser aplicados a:

$$L_{MVR}(\boldsymbol{\theta}, \boldsymbol{\lambda}; \mathbf{y}) = L_{MVR}(\boldsymbol{\theta}, \boldsymbol{\lambda}; \mathbf{y}^*) \prod_{i=1}^{N} \left|\boldsymbol{\Lambda}_i^{-1/2}\right|.$$

2.7.2 Decomposição da Matriz de Variância-Covariância $\boldsymbol{\Lambda}_i$

Podemos sempre decompor as matrizes $\boldsymbol{\Lambda}_i$ em produtos de matrizes mais simples:

$$\boldsymbol{\Lambda}_i = \mathbf{W}_i\mathbf{C}_i\mathbf{W}_i,$$

sendo $\mathbf{W}_i$ uma matriz diagonal e $\mathbf{C}_i$ uma matriz de correlação, isto é, uma matriz definida positiva onde todos os elementos da diagonal principal são iguais a 1. Uma vez que, ao multiplicar a matriz $\mathbf{W}_i$ por -1 (escalar), pode-se obter a mesma decomposição, não podemos afirmar que esta é única (Pinheiro & Bates, 2000). Para garantir a sua unicidade, precisamos de impor a positividade de todos dos elementos da diagonal.

É fácil verificar que

$$\mathrm{Var}(e_{it}) = \sigma^2 \left[\mathbf{W}_i\right]_{tt} \underbrace{\left[\mathbf{C}_i\right]_{tt}}_{=1} \left[\mathbf{W}_i\right]_{tt} = \sigma^2 \left[\mathbf{W}_i\right]_{tt}^2,$$

e

$$\text{Corr}(e_{it}, e_{it'}) = [\mathbf{C}_i]_{tt'},$$

de modo que, podemos afirmar que a matriz $\mathbf{W}_i$ descreve a variância e $\mathbf{C}_i$ a correlação dos erros aleatórios para o mesmo indivíduo. A decomposição de $\mathbf{\Lambda}_i$ nestas duas componentes, nomeadamente, a estrutura de variância e a estrutura de correlação, é conveniente quer a nível teórico, quer a nível computacional. É esta decomposição que nos vai permitir modelar as duas estruturas separadamente e combiná-las num modelo mais flexível (MLMG).

2.7.3 Funções Variância para a Modelação da Heterocedasticidade

As funções variância são usadas para modelar a estrutura de variância dos erros aleatórios individuais (apresentada na secção anterior) usando covariáveis.

Estas funções têm sido estudadas em detalhe no contexto do MLM por Davidian & Giltinian (1995) e, é de acordo com a parametrização proposta por estes autores que vamos definir a função variância dos erros individuais, $\mathbf{e}_i$, associados ao modelo (2.12):

$$\text{Var}(e_{it}|\mathbf{b}_i) = \sigma^2 g\left(\mu_{it}, \mathbf{v}_{it}, \boldsymbol{\delta}\right), \quad i = 1, \ldots, N, \quad t = 1, \ldots, n, \tag{2.13}$$

onde $\mu_{it} = \text{E}(e_{it}|\mathbf{b}_i))$, $\mathbf{v}_{it}$ é o vetor de covariáveis da variância, $\boldsymbol{\delta}$ é o vetor de parâmetros da covariância e $g(.)$ é a função variância, contínua em $\boldsymbol{\delta}$.

Por exemplo, se houver indícios de que a variabilidade inter-indivíduos aumenta com a potência de um valor absoluto de uma covariável v_{it}, então, podemos escrever:

$$\text{Var}(e_{it}|\mathbf{b}_i) = \sigma^2 \left|v_{it}\right|^{2\delta}. \tag{2.14}$$

A função varância neste caso é $g(x, y) = |x|^y$ e a covariável v_{it} pode ser o valor esperado u_{it}. Tal como este exemplo, poderíamos definir outros exemplos de funções variância com a função exponencial, logarítmica, ou até combinações destas, desde que a função escolhida reflita a variabilidade das medições realizadas ao mesmo indivíduo.

A formulação (2.14) é bastante flexivel e intuitiva, uma vez que permite que a variância inter-indivíduos dependa dos efeitos fixos, $\boldsymbol{\beta}$, e dos efeitos aleatórios, $\mathbf{b}_i$, através dos valores esperados μ_{it}. Contudo, apresenta algumas dificuldades teóricas e computacionais, à medida em que os erros individuais e os efeitos aleatórios deixam de ser independentes. É fácil de verificar que, partindo do pressuposto que $\text{E}(e_{it}|\mathbf{b}_i))$, $\text{Var}(e_{it}) = \text{E}[\text{Var}(e_{it}|\mathbf{b}_i)]$, pelo que a dependência dos erros aleatórios para o mesmo indivíduo, em relação aos efeitos aleatórios (não observados), pode ser evitada integrando em ordem aos efeitos aleatórios. Uma vez que a função variância g, em geral, não é linear em $\mathbf{b}_i$, a integração em ordem aos efeitos aleatórios pode ser complicada do ponto de vista computacional. Vamos, então, proceder de acordo com Davidian & Giltinian (1995) e usar uma aproximação do modelo em que os valores esperados μ_{it} são substituídos pelos respetivos BLUP, $\widehat{\mu_{it}} = \mathbf{x}_{it}^\top \boldsymbol{\beta} + \mathbf{z}_{it}^\top \widehat{\mathbf{b}}_i$, onde $\mathbf{x}_{it}$ e $\mathbf{z}_{it}$ denotam as t-ésimas linhas de $\mathbf{X}_i$ e $\mathbf{Z}_i$, respetivamente:

$$\text{Var}(e_{it}) \approx \sigma^2 g^2\left(\widehat{\mu_{it}}, \mathbf{v}_{it}, \boldsymbol{\delta}\right), \quad i = 1, \ldots, N, \quad t = 1, \ldots, n. \tag{2.15}$$

Segundo Pinheiro & Bates (2000), com esta aproximação, os erros aleatórios e os efeitos aleatórios deixam de estar correlacionados, como no modelo (2.12), sendo que os resultados obtidos na secção 2.7.1 continuam a ser válidos. É de notar que, se o modelo de variância condicional (2.13) não depender de μ_{it}, então não é necessário recorrer a uma aproximação, pois o modelo definido em (2.15) fornece a variância marginal exata.

A biblioteca *nlme* do R fornece um conjunto de funções variância, que são usadas para a modelação da heterocedasticidade, podendo ainda optar-se por usar uma combinação dessas mesmas funções. Na tabela seguinte são dadas algumas das funções variância implementadas no R:

Tabela 2.1: Funções variância para a modelação da heterocedasticidade

Descrição da Classe	$\mathrm{Var}(e_{it})$
Variância fixa	$\sigma^2 v_{it}$
Variâncias diferentes por estrato	$\sigma^2 \delta^2_{s_{it}}$
Potência de uma covariável	$\sigma^2 \lvert v_{it}\rvert^{2\delta}$
Exponencial de uma covariável	$\sigma^2 \exp(2\delta v_{it})$
Constante + Potência de uma covariável	$\sigma^2(\delta_1 + \lvert v_{it}\rvert^{\delta_2})^2$

v_{it} - covariável; s_{it} - variável de estratificação; $\delta_1 > 0$

Quando pretendemos comparar um modelo em que se assume a homocedasticidade com um modelo qualquer onde se assume a heterocedasticidade, usamos o teste de razão de verosimilhanças apresentado na secção 2.5.1, uma vez que os modelos têm a mesma estrutura fixa. Para a comparação de modelos com estruras fixas diferentes, tal como já foi referido anteriormente, usamos os critérios de informação AIC ou BIC.

2.7.4 Estruturas de Correlação para a Modelação da Dependência

As estruturas de correlação são usadas para modelar a dependência entre as observações. No contexto do MLM (básico e geral), são usadas para modelar a dependência dos erros aleatórios para medições realizadas ao mesmo indivíduo. Historicamente, as estruturas de correlação foram desenvolvidas para dois conjuntos principais de dados: dados temporais, onde as observações estão indexadas por uma variável temporal (unidimensional), e dados espaciais, referente a observações indexadas por um vetor espaço (bidimensional). Uma vez que os nossos dados se inserem no primeiro conjunto de dados, nesta secção apenas iremos abordar as estruturas de correlação serial consideradas mais importantes para o nosso estudo, sendo que podemos encontrar outros exemplos de estruturas de correlação serial e espacial em Pinheiro & Bates (2000).

De forma a estabelecer um quadro geral para as estruturas de correlação, vamos assumir que os erros aleatórios para o mesmo indivíduo, e_{it}, estão associados a vetores posição, $\mathbf{p}_{it}$. Para dados longitudinais, os vetores posição são, em geral, escalares inteiros. As estruturas de correlação aqui consideradas são supostas *isotrópicas* (Cressie, 1993), isto é, a correlação entre dois erros, e_{it} e $e_{it'}$, depende dos dos respetivos vetores posição, $\mathbf{p}_{it}$ e $\mathbf{p}_{it'}$, apenas através da distância entre eles, $\mathrm{dist}(\mathbf{p}_{it}, \mathbf{p}_{it'})$, e não dos valores particulares que assumem.

A estrutura de correlação geral para os erros aleatórios intra-indivíduos é dada por:

$$\mathrm{Corr}(e_{it}, e_{it'}) = h\left[\mathrm{dist}(\mathbf{p}_{it}, \mathbf{p}_{it'}), \boldsymbol{\rho}\right], \tag{2.16}$$

onde $\boldsymbol{\rho}$ é um vetor de parâmetros de correlação e $h(.)$ é uma função de correlação, que assume valores no intervalo $[-1, 1]$, contínua em $\boldsymbol{\rho}$ e tal que $h(0, \boldsymbol{\rho}) = 1$ (Pinheiro & Bates, 2000). Ou seja, quanto mais próximos estiverem os vetores posição de dois erros aleatórios (para o i-ésimo indivíduo), maior a sua dependência.

Estruturas de Correlação Serial

As estruturas de correlação serial são usadas para modelar dados observados sequencialmente ao longo do tempo e indexadas por um vetor posição unidimensional.

Simplificamos a hipótese de isotropia e assumimos que o modelo de correlação serial depende das posições, p_{it} e $p_{it'}$, apenas pela sua diferença absoluta. Podemos, então, reescrever (2.16) como:

$$\text{Corr}(e_{it}, e_{it'}) = h\left(\left|p_{it} - p_{it'}\right|, \boldsymbol{\rho}\right).$$

No contexto de dados indexados por uma variável temporal, a função de correlação, $h(.)$, é denominada função de autocorrelação. A **função de autocorrelação empírica**, descrita por Box *et al.* (1994), é uma estimativa não paramétrica da função de autocorrelação que fornece uma ferramenta útil para verificar a correlação serial em dados de séries temporais. Sejam

$$r_{it} = \frac{(y_{it} - \widehat{y_{it}})}{\widehat{\sigma_{it}}},$$

os resíduos padronizados do modelo linear misto ajustado, com $\sigma_{it}^2 = \text{Var}(e_{it})$. A função de autocorrelação empírica no espaçamento (*lag*) l é definida por:

$$\widehat{\rho}(l) = \frac{\sum_{i=1}^{N}\sum_{i=1}^{n-l} r_{it}r_{i(t+l)}/n(l)}{\sum_{i=1}^{N}\sum_{i=1}^{n} r^2it/n(0)},$$

onde $n(l)$ é o número de pares de resíduos utilizados no somatório do numerador de $\widehat{\rho}(l)$.

Em geral, as estruturas de correlação serial exigem que os dados sejam observados em instantes de tempo discretos, e não são facilmente generalizadas para vetores posição contínuos.

De seguida, descrevemos algumas das estruturas de correlação serial mais usadas, todas as quais estão implementadas na biblioteca *nlme* do R.

- **Simetria Composta**

Esta é a estrutura de correlação serial mais simples, onde se assume que todos os erros aleatórios pertencentes ao mesmo indivíduo têm igual correlação. O modelo de correlação correspondente é:

$$\text{Corr}(e_{it}, e_{it'}) = \rho, \quad \forall t \neq t', \quad h(k, \rho) = \rho, \quad k = 1, 2, \ldots,$$

onde o único parâmetro de correlação, ρ, é referido como como coeficiente de correlação intra-classe.

O modelo de correlação de simetria composta tende a ser muito simplista para situações práticas com dados de séries temporais, uma vez que é mais realista assumir um modelo em que a correlação entre duas observações diminui, em valor absoluto, com a distância. No entanto, é

útil para aplicações onde todas as observações para o mesmo indivíduo são recolhidas no mesmo instante de tempo.

- **Geral**

Esta estrutura representa o outro extremo, em termos de complexidade, da estrutura de simetria composta. Cada correlação entre as observações é dada por um parâmetro diferente, correspondendo à função de correlação:

$$h(k, \rho) = \rho, \quad k = 1, 2, \ldots$$

Esta estrutura apenas é útil em casos onde o número de observações por indivíduo (n) é pequeno, uma vez que o número de parâmetros aumenta quadraticamente com n.

- **Modelos auto-regressivos**

Os modelos auto-regressivos expressam uma determinada observação como a soma de uma função linear das observações anteriores com um termo de ruído homocedástico, a_t, centrado em zero ($\mathrm{E}(a_t) = 0$) e suposto independente das observações anteriores:

$$e_t = \phi_1 e_{t-1} + \cdots + \phi_p e_{t-p} + a_t.$$

O valor p, referente ao número de observações anteriores incluídas no modelo linear é designado por ordem do modelo auto-regressivo e denota-se AR(p).

O modelo **auto-regressivo de ordem 1**, AR(1), é o mais simples e um dos mais utilizados. A respetiva função de correlação diminui exponencialmente, em valor absoluto, com o espaçamento l e é dada por:

$$h(k, \phi) = \phi^k, \quad k = 0, 1, \ldots$$

O único parâmetro de correlação, ϕ, representa a correlação $l - 1$ e toma valores entre -1 e 1. Neste modelo, os erros no t-ésimo instante temporal são modelados como uma função do erro no instante $t - 1$, juntamente com o ruído:

$$e_t = \phi e_{t-1} + a_t, \quad |\phi| < 1.$$

O modelo AR(1) é um dos poucos modelos de correlação serial que podem ser generalizados para medições em tempo contínuo. Definimos a função de correlação AR(1) para tempo contínuo, denotada CAR(1), como:

$$h(s, \phi) = \phi^s, \quad s \geq 0, \quad \phi \geq 0.$$

Capítulo 3

Análise Estatística

A fim de ilustrar as diferentes metodologias mencionadas no capítulo anterior e derivar a relação linear que melhor traduz a taxa de incidência de TB nas Áreas Metropolitanas de Lisboa e do Porto, neste capítulo iremos descrever a análise estatística realizada e os respetivos resultados obtidos.

Como já foi referido anteriormente, toda a análise estatística foi implementada na linguagem R.

3.1 Base de Dados e Análise Descritiva

Nesta secção apresentamos uma descrição da base de dados (BD) em causa, assim como uma análise descritiva da mesma.

A BD contém dados estatísticos referentes a cada um dos 35 municípios considerados (18 da AML e 17 da AMP) para os anos 2010, 2011, 2012, 2013 e 2014.

Os dados relativos à variável resposta (taxa de incidência de TB, denominada TaxaTB) e à covariável "Número de casos de VIH por 100000 habitantes" foram fornecidos pela Prof. Raquel Duarte[1] e recolhidos das bases de dados do SVIG-TB e do INSA, respetivamente. Os dados relativos às restantes 26 variáveis explicativas foram recolhidos da base de dados do INE. Na tabela 3.1 podemos observar as covariáveis consideradas no estudo, assim como as respetivas denominações que iremos usar a partir de agora.

Tabela 3.1: Sumário das covariáveis consideradas e respetivas denominações

Covariável	**Denominação**
Número de casos de VIH por 10^5 habitantes	VIH
Percentagem da população total residente entre os 0 e os 14 anos	Pop0_14
Percentagem da população total residente entre os 15 e os 24 anos	Pop15_24
Percentagem da população total residente entre os 25 e os 64 anos	Pop25_64
Percentagem da população total residente entre os 65 e os 74 anos	Pop65_74
Percentagem da população total residente com idade $\geq$ 75 anos	Pop75

[1]Programa Nacional para a Tuberculose; Centro de Referência Nacional para a Tuberculose Multi-resistente; UGI Torax, Centro Hospitalar de Vila Nova de Gaia/Espinho; Faculdade de Medicina, Universidade do Porto; Instituto de Saúde Pública, Universidade do Porto.

Tabela 3.1: Sumário das covariáveis consideradas e respetivas denominações

Covariável	**Denominação**
Percentagem da população total residente do sexo masculino	PopMasc
Taxa de criminalidade	TaxaCriminalidade
Número de crimes contra a integridade física por 10^3 habitantes	IntFisica
Número de crimes de furto de veículo e em veículo motorizado por 10^3 habitantes	FurtoVeic
Número de crimes por condução de veículo com taxa de álcool ≥ 1.2 g/l por 10^3 habitantes	CondAlcool
Número de crimes por condução sem habilitação legar por 10^3 habitantes	SemCarta
Número de crimes contra o património por 10^3 habitantes	CrimPatr
Taxa bruta de mortalidade	TaxaMortalidade
Número de mortes por tumores (neoplasmas) por 10^3 habitantes	MortTumores
Número de mortes por dependência de drogas por 10^6 habitantes	MortDrogas
Número de mortes por doenças do aparelho circulatório por 10^4 habitantes	MortCirculatorio
Número de mortes por doenças isquémicas do coração por 10^4 habitantes	MortCoracao
Número de mortes por outras doenças cardíacas[2] por 10^4 habitantes	MortOutrasCoracao
Número de mortes por doenças cérebro-vasculares por 10^4 habitantes	MortCerebro
Número de mortes por doenças do aparelho respiratório por 10^4 habitantes	MortRespiratorio
Número de médicos por 10^3 habitantes	Medicos
Número de enfermeiros por 10^3 habitantes	Enfermeiros
Densidade populacional	DensPop
Número de beneficiários do RSI[3] por 10^2 habitantes em idade ativa[4]	BenRSI
População estrangeira que solicitou estatuto de residente por 10^4 habitantes	PopEstrangeira

É de referir que os dados do INSA e do INE (com exceção das variáveis TaxaCriminalidade, TaxaMortalidade, DensPop e BenRSI) foram alterados para que pudessem traduzir uma proporção da população total residente, sendo que originalmente eram contagens.

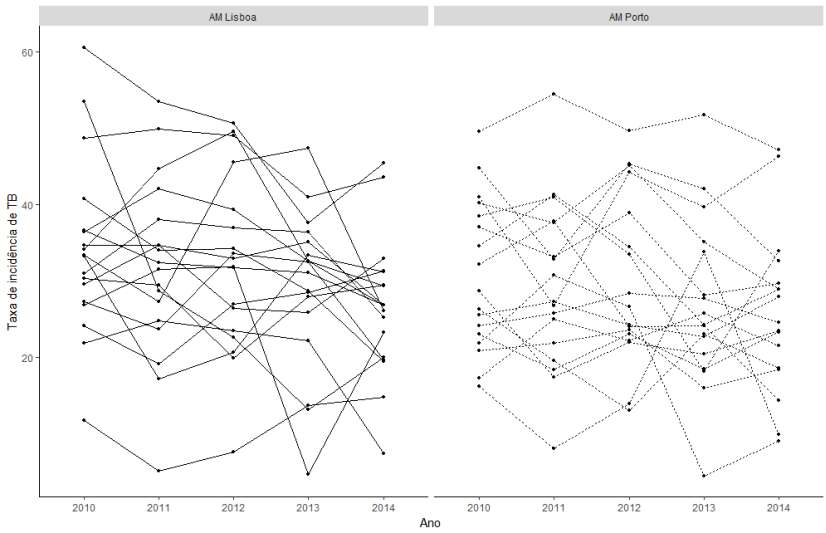
Uma vez que a BD contempla 26 variáveis explicativas, e que seria dispendioso expor a análise univariada realizada a cada uma das variáveis da BD, optamos por apresentar apenas a análise

[2]Exceto transtornos valvulares não-reumáticos e doenças valvulares.

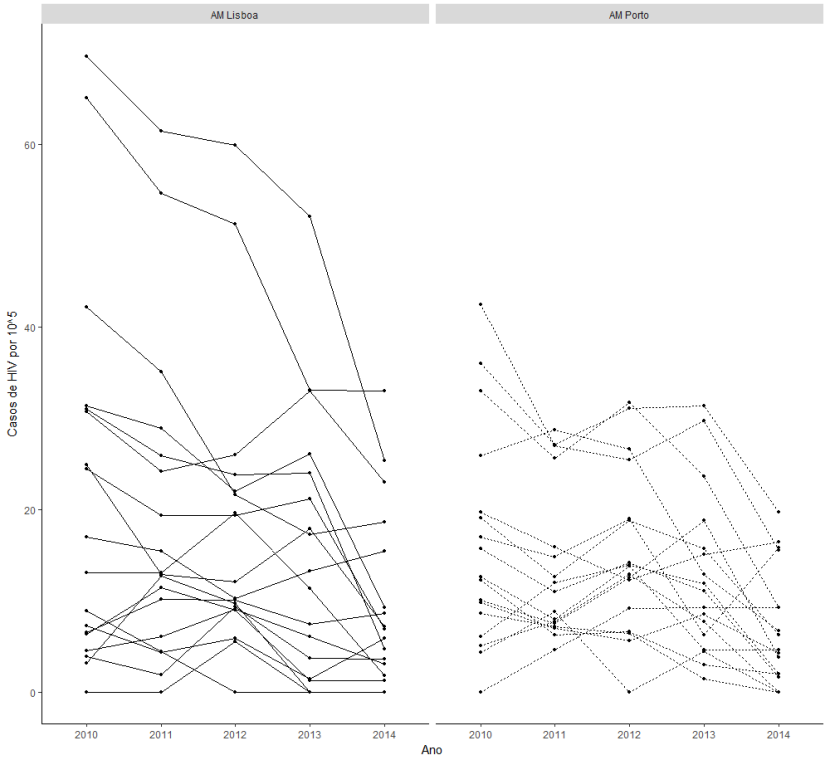
[3]Rendimento Social de Inserção.

[4]População com idade mínima de 15 anos que, no período de referência, constituía a mão de obra disponível para a produção de bens e serviços que entram no circuito económico (população empregada e desempregada).

das variáveis que se irão revelar significativas para o estudo. Assim, a figura 3.1 representa o perfil temporal de cada município relativamente às variáveis TaxaTB (resposta), VIH, BenRSI e PopEstrangeira (covariáveis), por área metropolitana.

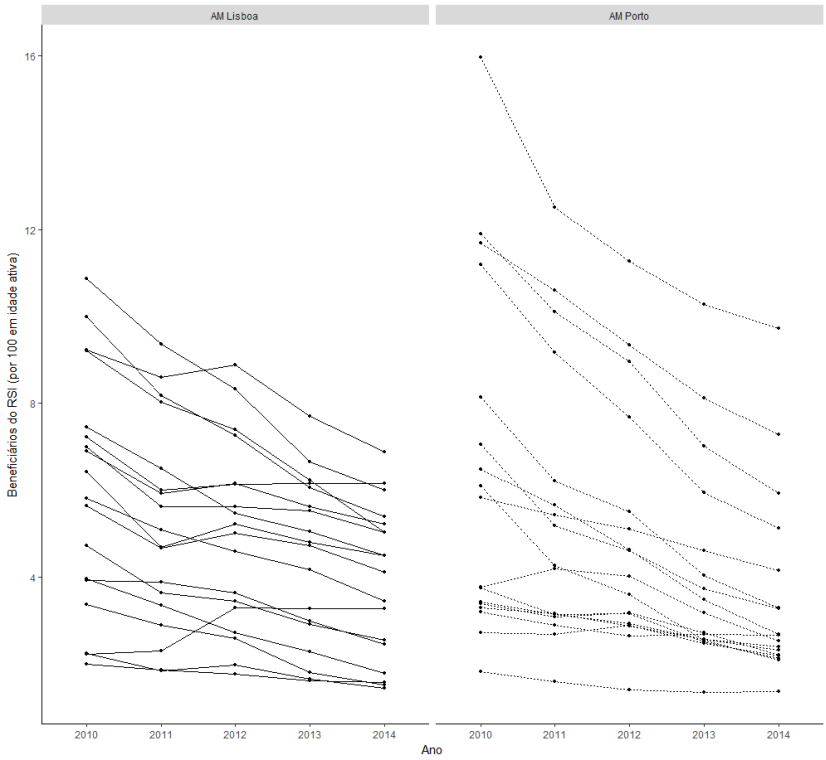


(a) TaxaTB

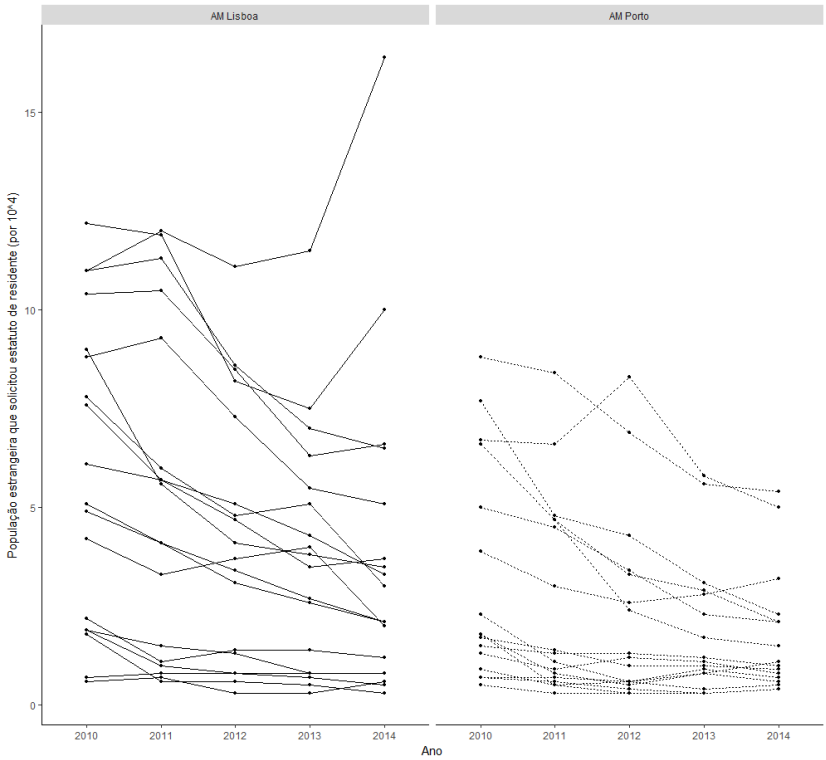


(b) VIH

Figura 3.1: Análise univariada das características da BD (parte 1)



(c) BenRSI



(d) PopEstrangeira

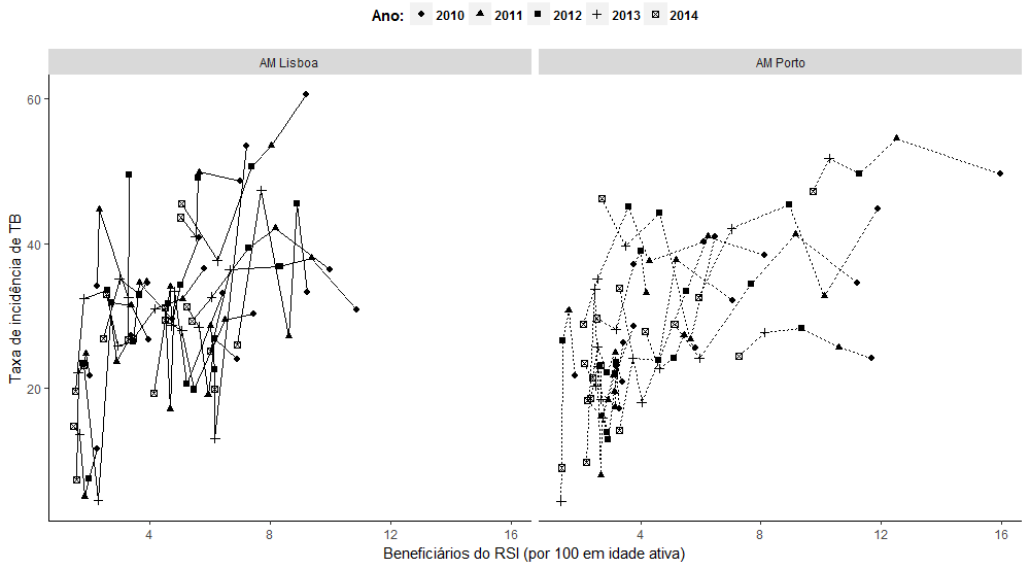
Figura 3.1: Análise univariada das características da BD (parte 2)

Podemos verificar que as variáveis VIH, BenRSI e PopEstrangeira parecem estar a diminuir com o tempo, tendência esta que não é tão evidente na variável resposta. Para além disso, as duas populações (AML e AMP) apresentam diferenças claras em relação às covariáveis consideradas:

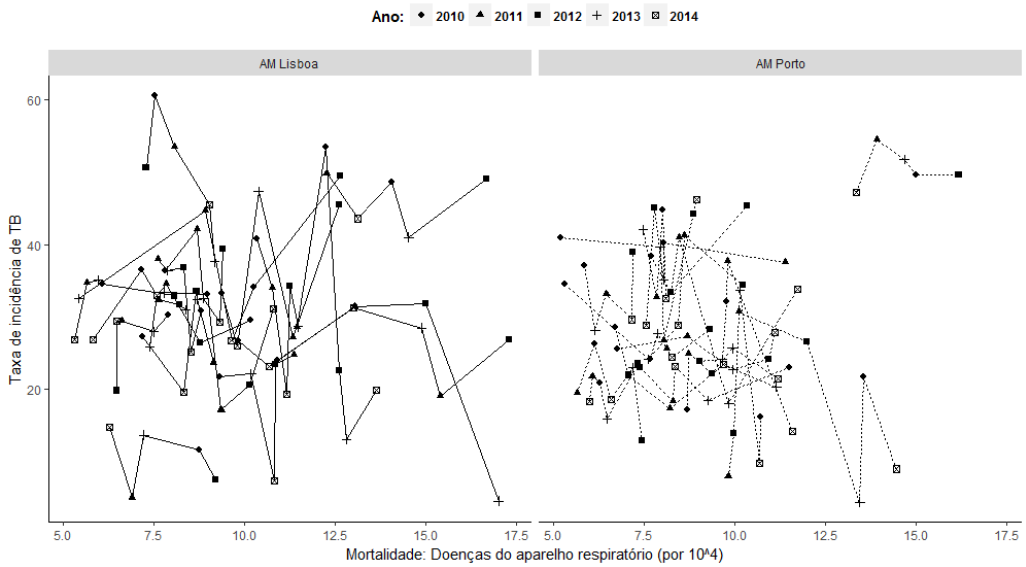
a AML tem municípios com valores de VIH e PopEstrangeira mais altos, enquanto que para a covariável BenRSI apresenta valores inferiores aos municípios da AMP. Mais uma vez, não existe tal discrepância na variável resposta, pelo que não há indícios de que os municípios de uma das áreas metropolitanas tenha taxa de incidência de TB mais alta.

3.2 Análise Exploratória

Ao longo da análise exploratória foi possível distinguir três tipos de situações: aquelas em que parece existir uma relação linear entre a variável resposta e a covariável em questão, aquelas em que não é possível discernir se tal relação existe, e ainda aquelas onde podemos verificar que não há uma relação linear.

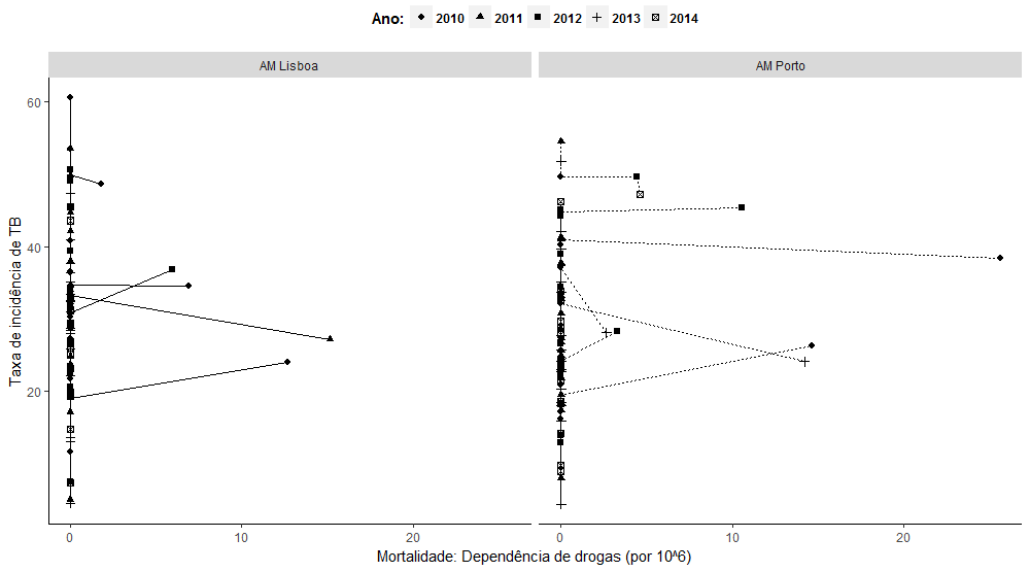


(a) Relação entre a TaxaTB e BenRSI



(b) Relação entre a TaxaTB e MortRespiratorio

Figura 3.2: Relações entre a variável resposta e algumas covariáveis (parte 1)



(c) Relação entre a TaxaTB e MortDrogas

Figura 3.2: Relações entre a variável resposta e algumas covariáveis (parte 2)

Na figura 3.2 podemos observar três exemplos destas situações: (a) representa um exemplo da primeira situação, onde podemos observar que taxa de incidência de TB parece crescer linearmente com a proporção de beneficiários de RSI; (b) corresponde a um exemplo onde não é explicita uma tendência linear que descreva o comportamento da incidência de TB através da proporção de óbitos por doenças do aparelho respiratório; (c) corresponde a um exemplo da última situação acima descrita, onde podemos verificar que não existe uma relação linear entre a variável resposta e a covariável em questão (por exemplo, os municípios com menor e maior taxa de incidência de TB têm proporções de mortes por dependência de drogas bastante semelhantes enquanto que, municípios com incidência de TB intermédios, têm valores mais altos).

O processo que levou à identificação da estrutura fixa do modelo final começou com uma análise de regressão simples (apenas com um preditor), com e sem a variável temporal no preditor linear, com e sem termos de interação e com efeito aleatório no termo constante. A escolha do modelo final seguiu o algoritmo de seleção direta: resumidamente, o preditor linear começou com uma única covariável, a mais significativa nas regressões individuais e, posteriormente, adicionamos uma covariável de cada vez, de acordo com sua significância estatística. A comparação de modelos com diferentes estruturas fixas foi feita com base nos critérios de informação descritos na secção 2.5.3.

A parte fixa do modelo é, então, composta pelas variáveis VIH, BenRSI e PopEstrangeira. A variável temporal, por si só, também não se revelou estatisticamente significativa. Para além das covariáveis do modelo final, as principais variáveis que se mostraram significativas na análise de regressão simples foram PopMasc, Medicos, Enfermeiros e DensPop.

Com o objetivo de identificar os efeitos aleatórios a considerar, foi feito um ajustamento linear individual (figura 3.3) para analisar quais os parâmetros que mais variam de município para município. Não sendo totalmente adequada, pelo reduzido número de observações relativamente ao número de parâmetros de cada modelo, esta metodologia permite avaliar a necessidade da inclusão de efeitos aleatórios.

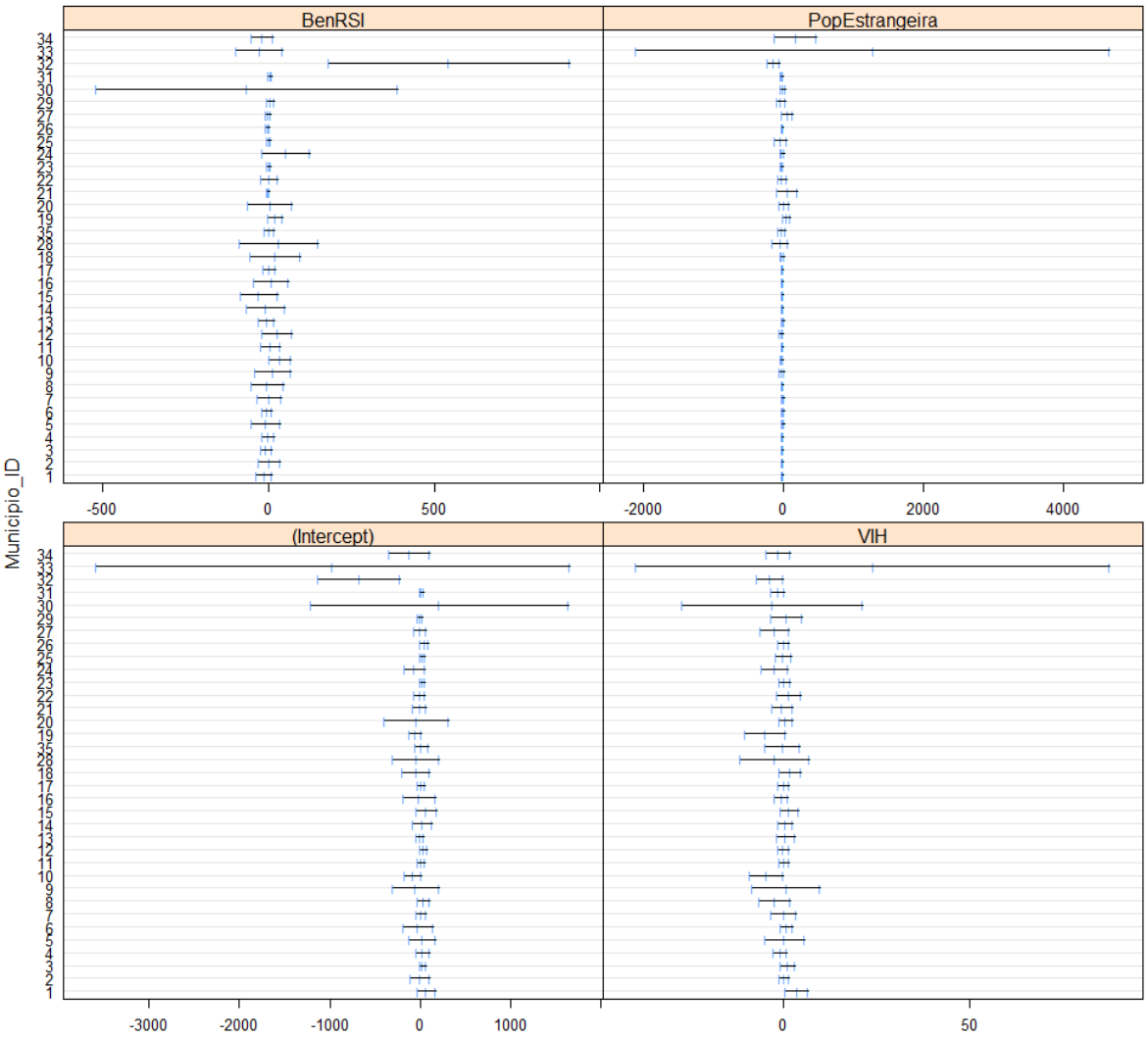


Figura 3.3: Estimativas dos intervalos de confiança para os parâmetros do ajustamento individual

Como podemos verificar, não há grande variabilidade nas estimativas do termo constante e dos restantes coeficientes para as variáveis VIH, BenRSI e VIH. Ainda assim, consideramos dois modelos (Mod_1 e Mod_2) com dois efeitos aleatórios no termo constante e na covariável BenRSI. Em ambos os modelos consideramos o fator de agrupamento (AM - área metropolitana) para verificar a existência de diferenças entre as duas populações (uma vez que este é invariante no tempo, apenas é necessário o indíce i). Esta escolha deve-se ao facto de, na análise de regressão univariada mencionada acima, o termo de interação ter sido estatisticamente significativo para a covariável BenRSI.

- **Mod$_1$**

$$Y_{it} = (\beta_0 + b_{0i}) + \beta_1 VIH_{it} + (\beta_2 + b_{2i}) BenRSI_{it} + \beta_3 PopEstrangeira_{it} + \beta_4 AM_i + e_{it},$$

- **$\mathbf{Mod_2}$**

$$Y_{it} = (\beta_0 + b_{0i}) + \beta_1 VIH_{it} + (\beta_2 + b_{2i}) BenRSI_{it} + \beta_3 PopEstrangeira_{it} + \beta_4 AM_i + \\ + \beta_5 (BenRSI_{it} \times AM_i) + e_{it},$$

com $i = 1, \ldots, 35$, $t = 2010, \ldots, 2014$ e onde Y_{it} é a variável aleatória que representa a taxa de incidência de TB para o i-ésimo indivíduo no t-ésimo instante de observação. AM_i é uma variável binária que assume o valor 0 se o i-ésimo indivíduo pertence à AML e o valor 1 caso contrário. Os parâmetros $\beta_0, \ldots, \beta_5$ representam os efeitos fixos do modelo, o vetor aleatório $\mathbf{b}_i = (b_{0i}, b_{2i})^\top$ representa os efeitos aleatórios e e_{it} designa o erro aleatório.

Tendo em conta que o valor-p da componente $BenRSI_{it} \times AM_i$ do Mod_2 é superior a 0.05, e que os respetivos valores de AIC e BIC são superiores aos do Mod_1, optamos por continuar com o Mod_1. Passamos então ao reajustamento do modelo selecionado usando o método MVR, e analisamos a possibilidade de haver apenas um efeito aleatório: ou no termo constante (Mod_3) ou na variável BenRSI (Mod_4).

Tabela 3.2: Comparação entre os modelos aninhados e o modelo geral

Modelo	AIC	BIC	logLIK	Teste	TRV	valor-p
Mod_1	1211.408	1239.631	-596.704			
Mod_3	1222.836	1244.787	-604.418	1 *vs* 3	15.428	< 0.01
Mod_4	1208.997	1230.947	-597.498	1 *vs* 4	1.588	0.452

TRV - Teste da Razão de Verosimilhanças

Ou seja, não rejeitamos a estrutura aleatória do Mod_4, sendo este o modelo com que continuamos. A equação do modelo é dada por (3.1) e o respetivo sumário encontra-se na tabela 3.3.

$$Y_{it} = (\beta_0 + b_{0i}) + \beta_1 VIH_{it} + \beta_2 BenRSI_{it} + \beta_3 PopEstrangeira_{it} + \beta_4 AM_i + e_{it}. \quad (3.1)$$

Tabela 3.3: Sumário do Mod_4

Efeitos Fixos					
Parâmetro	Estimativa MVR	SE	GL	*estatística t*	*valor-p*
β_0	13.853	2.321	137	5.969	< 0.01
β_1	0.193	0.073	137	2.637	< 0.01
β_2	1.435	0.338	137	4.242	< 0.01
β_3	0.950	0.362	137	2.626	< 0.01
β_4	5.331	2.534	33	2.104	0.043
Efeitos Aleatórios					
	Termo constante	Resíduo			
SD	4.706	6.580			

SE - Erro Padrão; GL - Graus de Liberdade; SD - Desvio Padrão

A variabilidade inter-indivíduos explicada pelo efeito aleatório na constante é de $\frac{4.71^2}{4.71^2+6.58^2} \times 100 \approx 33.88\%$, que é bastante alta se considerarmos que a análise gráfica da estimativa dos

intervalos de confiança para os parâmetros do ajustamento individual (figura 3.3) não apontava para a necessidade da inclusão de efeitos aleatórios.

É de notar que, uma vez que temos apenas um efeito aleatório, não é necessário proceder à modelação da estrutura de variância dos efeitos aleatórios (matriz **D**).

Diagnóstico e Modelação da Heterocedasticidade e da Dependência

O gráfico dos resíduos padronizados *vs.* os valores ajustados para o Mod_4 é apresentado na figura 3.4, onde podemos observar uma ligeira diminuição da variabilidade com os valores ajustados.

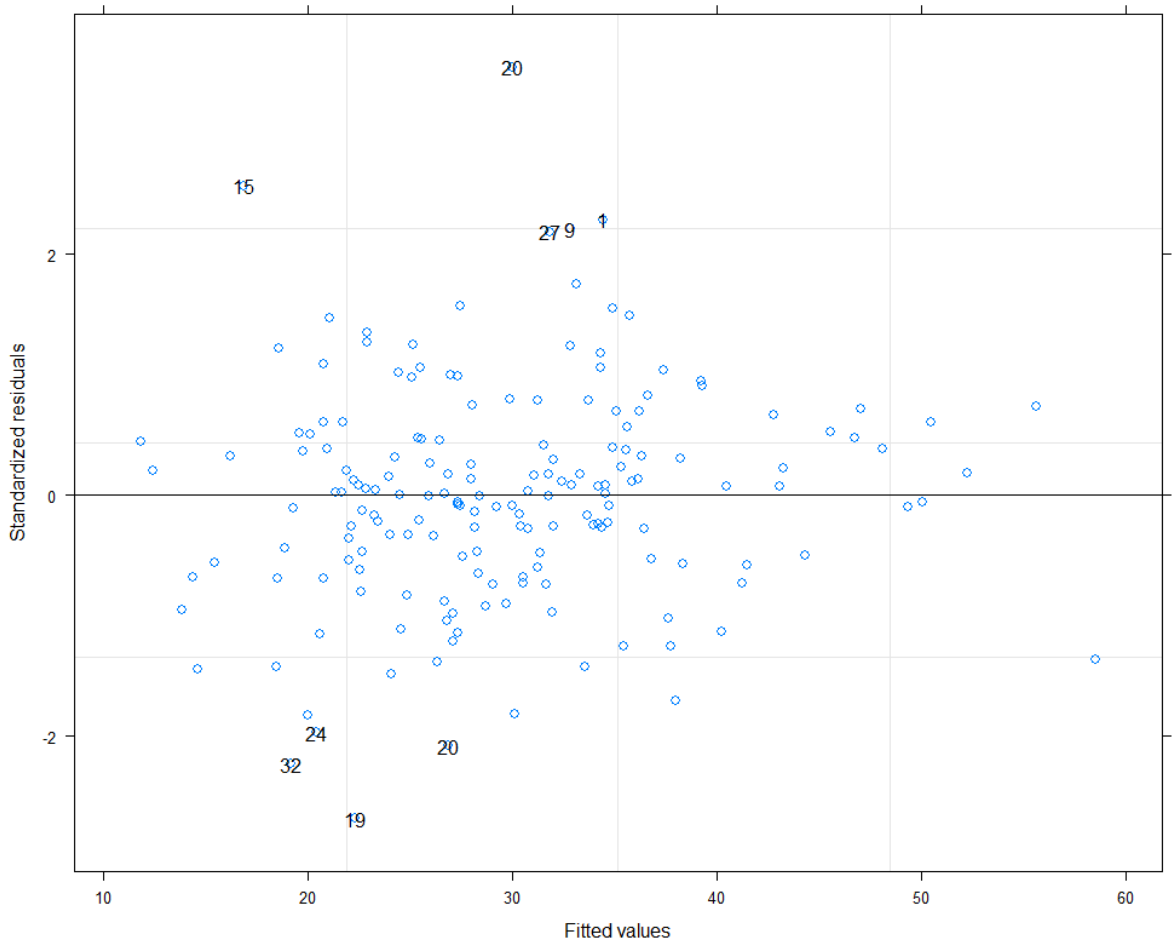


Figura 3.4: Resíduos padronizados *vs.* os valores ajustados para o Mod_4

Podemos visualizar a mesma representação gráfica na figura 3.5, agrupada por área metropolitana. Esta, sugere uma eventual heterocedasticidade entre áreas metropolitanas, uma vez que a variabilidade parece ser um pouco maior na AMP.

Para modelar a heterocedasticidade observada nos gráficos anteriores, usamos as funções variância: potência de uma covariável, $\text{Var}(e_{it}) = \sigma^2|v_{it}|^{2\sigma}$ (Mod_5), sendo igualmente plausível variâncias diferentes por estrato, $\text{Var}(e_{it}) = \sigma^2\delta^2_{S_{it}}$ (Mod_6). Com base no teste da razão de verosimilhanças apresentado na tabela 3.4, onde comparamos cada um destes modelos com o

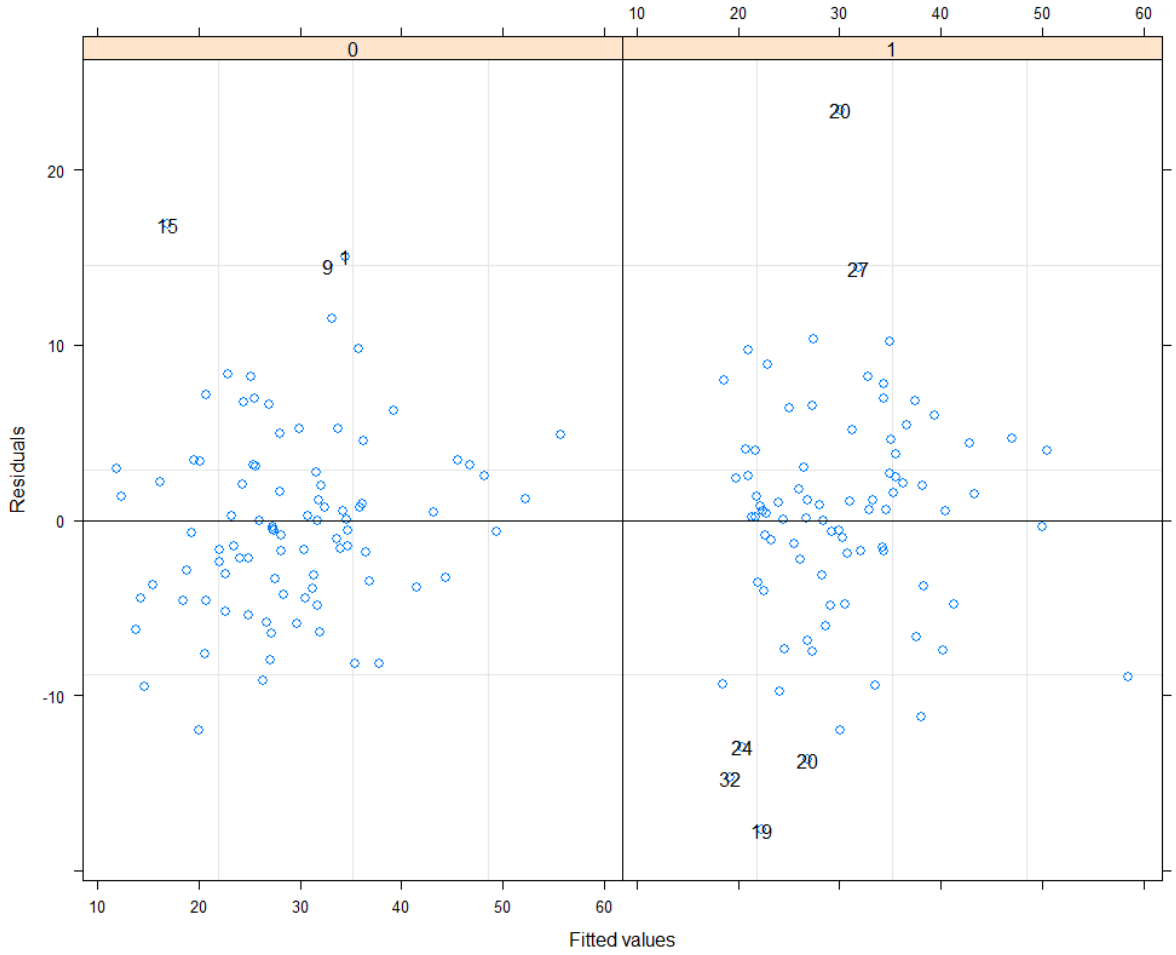


Figura 3.5: Resíduos padronizados *vs.* os valores ajustados por área metropolitana para o Mod_4

Mod_4, verificamos que os valores-p não indicam evidência de heterocedasticidade, pelo que os modelos aninhados são rejeitados.

Tabela 3.4: Modelação da heterocedasticidade - comparação de modelos

Modelo	AIC	BIC	logLIK	Teste	TRV	valor-p
Mod_4	1208.997	1230.947	−597.498			
Mod_5	1209.000	1234.086	−596.500	4 *vs* 5	1.997	0.158
Mod_6	1207.872	1232.959	−595.936	4 *vs* 6	3.124	0.077

Relativamente à modelação da dependência dos erros aleatórios, procedemos à comparação de modelos cujas estruturas de correlação serial foram estudadas na secção 2.7.4. Temos, então, a estrutura de simetria composta (Mod_7), a geral (Mod_8) e o modelo auto-regressivo de ordem 1 (Mod_9). Na tabela 3.5, podemos observar o resultado do teste da razão de verosimilhanças realizado para comparar estes três modelos com o Mod_4.

Podemos verificar que os valores-p dos TRV realizados para os Mod_8 e Mod_9 indicam forte evidência de dependência, pelo que o Mod_4 é rejeitado em ambos. Segue-se a comparação entre o Mod_8 e o Mod_9, feita com base nos critérios de informação AIC e BIC, uma vez que estes não estão aninhados. Dessa comparação, selecionamos o Mod_9 pois, apesar de ter um valor de AIC

Tabela 3.5: Modelação da dependência - comparação de modelos

Modelo	AIC	BIC	logLIK	Teste	TRV	valor-p
Mod_4	1208.997	1230.947	-597.498			
Mod_7	1210.997	1236.083	-597.498	4 *vs* 7	0	1
Mod_8	1200.074	1253.382	-583.037	4 *vs* 8	28.923	< 0.01
Mod_9	1201.157	1226.243	-592.578	4 *vs* 9	9.840	< 0.01

mais elevado, a diferença entre este e o Mod_8 não é significativa (≈ 1), sendo que tem um valor de BIC mais baixo, com maior discrepância (≈ 27).

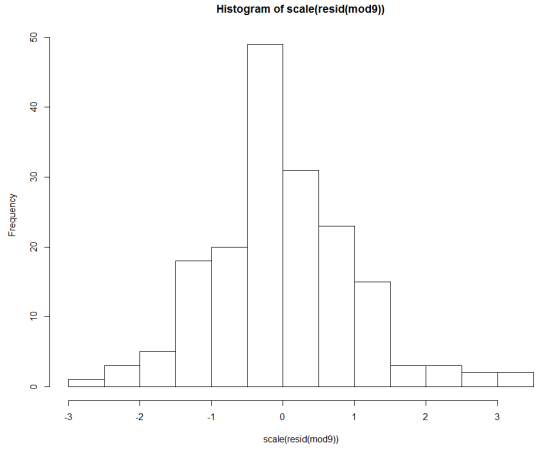
3.3 Resultados

O modelo final é descrito pela equação (3.1), uma vez que a estrutura fixa e aleatória é a mesma que a do Mod_4. Na tabela 3.6 apresentamos as estimativas obtidas para as componentes fixas e aleatórias do modelo. A variância residual foi de $7.3^2 = 53.29$ e o coeficiente de correlação, correspondente a uma estrutura do tipo autoregressiva de ordem 1, foi de 0.350.

Tabela 3.6: Sumário do modelo final

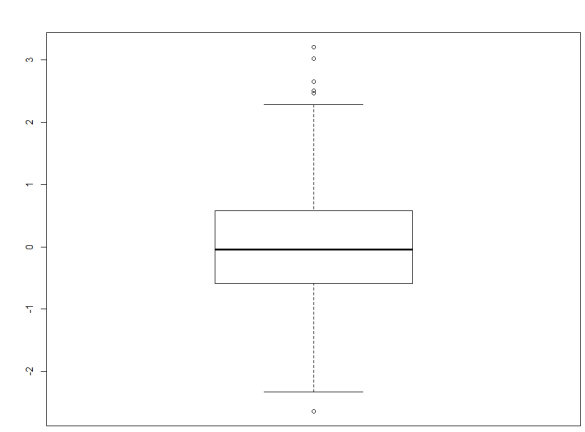
Efeitos Fixos					
Parâmetro	Estimativa MVR	SE	GL	estatística t	valor-p
β_0	12.893	2.398	137	5.376	< 0.01
β_1	0.180	0.074	137	2.442	0.016
β_2	1.512	0.349	137	4.329	< 0.01
β_3	1.098	0.377	137	2.917	< 0.01
β_4	5.648	2.555	33	2.211	0.034
Efeitos Aleatórios					
	Termo constante				
SD	3.330				

O passo seguinte é analisar se o modelo verifica os pressupostos requeridos, através da figura 3.6.

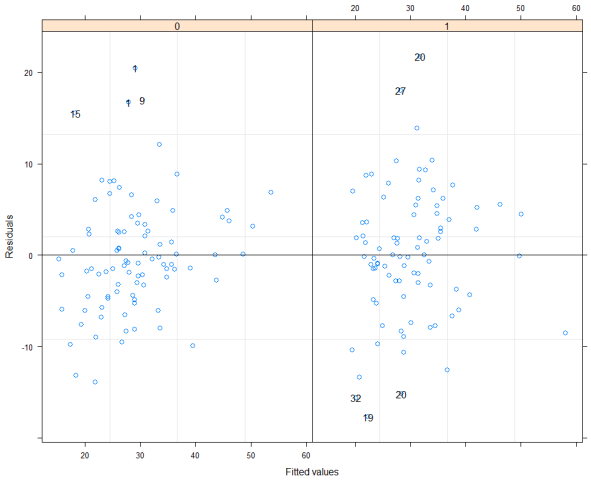


(a) Histograma dos resíduos

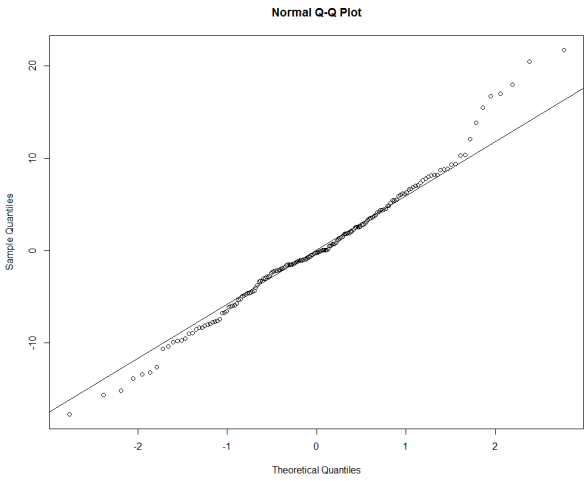
Figura 3.6: Gráficos de diagnóstico do modelo final (parte 1)



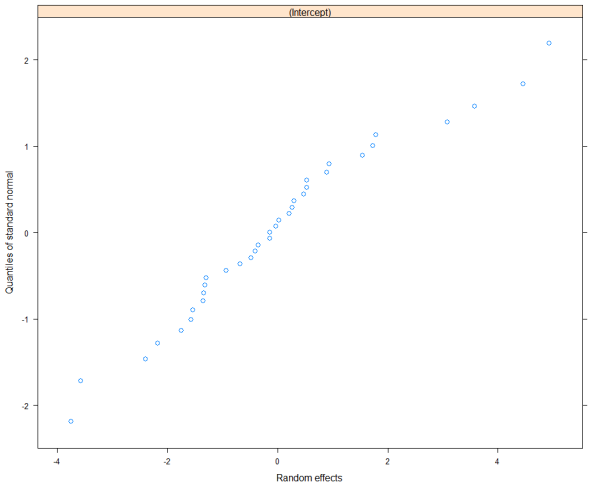
(b) Boxplot dos resíduos



(c) Resíduos *vs.* valores ajustados por AM



(d) Resíduos *vs.* Gráfico QQ dos resíduos



(e) Gráfico QQ dos efeitos aleatórios

Figura 3.6: Gráficos de diagnóstico do modelo final (parte 2)

Apesar do gráfico (b) revelar a existência de alguns outliers, o histograma (a) apresenta uma simetria razoável. Os poucos outliers existentes não parecem estar a interferir grandemente com a normalidade dos resíduos, pelo gráfico dos quantis (d), e portanto não procedemos à remoção destas observações da BD. Apesar de, no modelo final, não ter sido alterada a estrutura de variância dos erros aleatórios, a heterocedasticidade entre áreas metropolitanas verificada na figura 3.5 já não é evidente em comparação com (c). O gráfico (e) é difícil de avaliar, uma vez que a BD apenas contempla 35 municípios. No entanto, não descartamos o pressuposto de normalidade dos efeitos aleatórios na constante. Podemos então considerar que as hipóteses subjacentes ao MLM são verificadas e afirmar que a evolução da incidência de TB é uma função linear das covariáveis consideradas. Também podemos afirmar que o aumento verificado na incidência de TB da AMP em relação à AML é estatisticamente significativo (valor-p= 0.034). Mais ainda, a taxa de incidência de TB aumenta linearmente com as variáveis VIH, BenRSI e PopEstrangeira. Isto é, quanto maior forem as proporções de casos de VIH, de beneficiários de RSI e de indivíduos estrangeiros a solicitarem estatuto de residência, maior será a incidência de TB. O aumento verificado na incidência de TB da AMP relativamente à AML diz-nos que, se a

população da AMP fosse igual à da AML em relação às covariáveis do estudo, então teria uma taxa de incidência mais alta.

A percentagem de variabilidade da resposta explicada pelo efeito aleatório diminuiu em relação ao Mod_4, ainda assim, foi de $\frac{3.33^2}{3.33^2+7.30^2} \times 100 \approx 17.22\%$.

Relativamente à matriz dos erros aleatórios, a estrutura de correlação intra-indivíduo (matriz $\mathbf{C}_i$) é explicada por um modelo auto-regressivo de ordem 1 com $\phi = 0.350$.

Na tabela 3.7 podemos observar as estimativas EBLUP dos efeitos aleatórios na constante.

Tabela 3.7: Estimativas EBLUP dos efeitos aleatórios

AML		AMP	
Município	BLUP	Município	BLUP
Alcochete	4.928	Arouca	-0.035
Almada	0.531	Espinho	-0.403
Amadora	3.079	Gondomar	-0.479
Barreiro	-0.139	Maia	-1.341
Cascais	-1.573	Matosinhos	0.295
Lisboa	0.530	Oliveira de Azeméis	-1.540
Loures	-0.354	Paredes	0.020
Mafra	-3.746	Porto	0.471
Moita	1.781	Póvoa de Varzim	4.456
Montijo	0.211	Santa Maria da Feira	-0.931
Odivelas	0.935	Santo Tirso	-0.684
Oeiras	0.266	São João da Madeira	-1.348
Palmela	-0.143	Trofa	1.728
Seixal	-2.177	Vale de Cambra	-1.745
Sesimbra	-2.402	Valongo	1.541
Setúbal	-1.299	Vila do Conde	3.573
Sintra	0.889	Vila Nova de Gaia	-3.577
Vila Franca de Xira	-1.319		

É de notar que se todos os municípios tivessem a mesma população em relação às variáveis VIH, BenRSI e PopEstrangeira, então a Póvoa de Varzim (respetivamente Vila Nova de Gaia) seria o município com maior (respetivamente menor) incidência de TB na AMP. Do mesmo modo, Alcochete (respetivamente Mafra) apresentaria a maior (respetivamente menor) taxa de incidência de TB na AML.

O modelo de regressão com as variáveis padronizadas, cujo sumário está descrito na tabela 3.8, evidencia o RSI como a variável explicativa mais influente na incidência de TB.

Tabela 3.8: Sumário do modelo final com variáveis padronizadas

Efeitos Fixos					
Parâmetro	Estimativa MVR	SE	GL	estatística t	valor-p
β_0	26.857	1.546	137	17.369	< 0.01
β_1	2.408	0.986	137	2.442	0.016
β_2	4.069	0.940	137	4.329	< 0.01
β_3	3.626	1.243	137	2.917	< 0.01
β_4	5.648	2.555	33	2.211	0.034

As figuras 3.7 e 3.8 representam, respetivamente, os gráficos dos valores observados *vs.* valores

ajustados e o ajustamento individual com base no Mod_9.

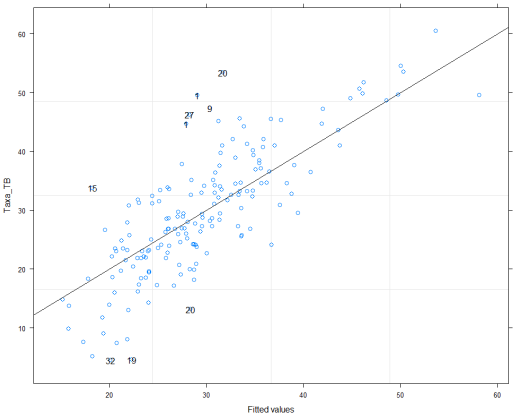


Figura 3.7: Valores observados *vs.* valores ajustados

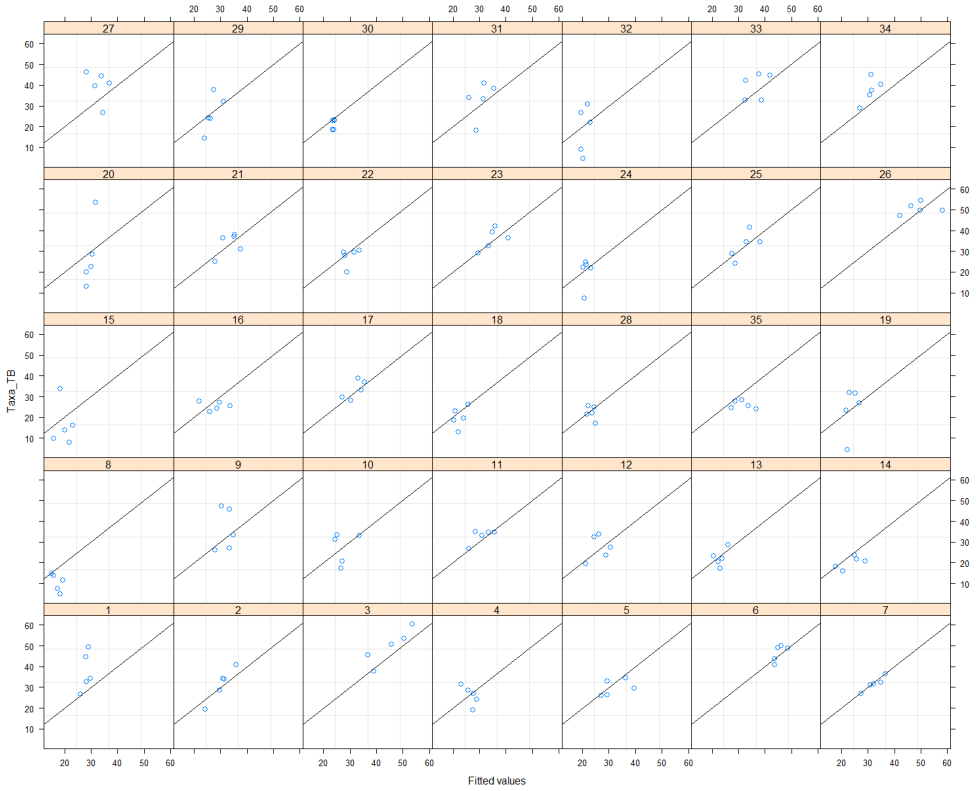


Figura 3.8: Ajustamento individual com base no Mod_9

Capítulo 4

Considerações Finais e Trabalho Futuro

O objetivo proposto pela Prof. Raquel Duarte, coordenadora do Programa Nacional para a Tuberculose, foi estudar a disseminação da tuberculose nas duas maiores áreas metropolitanas de Portugal, através da identificação dos principais fatores de risco.

Dada a estrutura longitudinal dos dados, e para ter em consideração a variação regional verificada em cada área metropolitana, usou-se o Modelo Linear Misto estudado no capítulo 2. Em primeiro lugar, foi conduzida uma análise de regressão simples que, através do algoritmo de seleção direta, permitiu identificar a melhor estrutura fixa. A comparação de modelos com diferentes estruturas fixas foi feita com base nos critérios de informação. De seguida, foi identificada a estrutura aleatória através das estimativas dos intervalos de confiança para os parâmetros do ajustamento individual. Posteriormente, realizou-se uma análise da heterocedasticidade e da dependência dos erros aleatórios, que resultou na modelação da estrutura de correlação dos mesmos segundo um modelo autoregressivo de ordem 1. A comparação entre modelos aninhados baseou-se em testes de razão de verosimihanças. Desta forma, observou-se uma associação significativa e positiva entre as variáveis VIH, RSI e PopEstrangeira e a taxa de incidência de TB. O modelo final mostrou que, entre estas covariáveis, o RSI (usado neste estudo como uma medida de pobreza) é o que tem maior peso no desenvolvimento da doença.

Os resultados obtidos ao longo deste trabalho são suportados por estudos prévios, sendo que já está bem estabelecida a associação entre as variáveis do modelo final e a incidência de TB. A pobreza é por si só uma variável complexa na medida em que pode ser associada a um amplo grupo de fatores sociais (como condições de habitação, desemprego, educação, etc) que atuam de forma sinérgica. Apesar da maioria dos estudos publicados até hoje terem em consideração diferentes medidas para avaliar esta variável, todos estes indicam uma forte associação entre a pobreza e o desenvolvimento da TB. [2, 5, 10] A imigração oriunda de países com alta incidência da doença também já foi descrita como um fator de risco, embora alguns estudos europeus apontem para a possibilidade de a imigração não ser um dos fatores que mais contribui para o aumento da incidência de TB da população nativa europeia. [5, 23, 25, 26, 33] Contudo, de acordo com dados da DGS, a incidência em cidadãos estrangeiros representou cerca de 15.6% do total de casos em 2012. [11] Neste trabalho, a imigração foi incluída através da variável PopEstrangeira. Para além dos fatores socio-económicos, outro fator de risco que se destacou foi a infeção por VIH. De acordo com o Centro Europeu de Prevenção e Controlo das Doenças, em 2015 Portugal fez parte do grupo de países europeus com maior proporção de casos de co-infeção VIH/TB (10 a

14.9%). [15] Segundo um estudo europeu publicado em 2011, referente ao período compreendido entre 2001 e 2008, a percentagem de co-infeção era de 14.6%, apesar de, em Portugal, o número de novos casos de infeção por VIH ter diminuído cerca de 73.5% entre 2000 e 2016. [13, 31]

O modelo final mostrou que, embora o perfil da taxa de incidência de TB varie de forma idêntica em ambas as áreas metropolitanas, se a população da AMP fosse a mesma que a da AML, relativamente às covariáveis VIH, RSI e PopEstrangeira, a incidência seria mais elevada na AMP. Êstes resultados levam-nos a concluir que pode haver fatores, para além daqueles considerados, que influenciem a taxa de incidência de TB na AMP. De acordo com estudos publicados, a população da AMP tem menor nível de escolaridade, maior taxa de desemprego e de superlotação habitacional, comparativamente com a AML. [5, 21, 33] Para além disso, parece haver uma dimensão histórico-social que influencia a taxa de incidência de TB a nível regional. No final do século XIX, 21.4% da população da AMP vivia em ilhas[1], que passaram a ser consideradas como focos de infeção, nomeadamente de TB. [24] Esta realidade habitacional persistiu até aos dias de hoje. Entre 1960 e 1980, a AMP enfrentou um crescimento populacional de, aproximadamente, 34%. [29] Consequentemente, houve um crescimento residencial sub-urbano que foi realizado através da expansão dos segmentos da população rural antiga, e não através da construção de novos bairros, mantendo assim o mesmo perfil arquitetónico [29] - facto que pode ter contríbuido para a tendência de maior incidência de TB na AMP. Também é possível explorar a possibilidade da heterogeneidade verificada entre as duas áreas metropolitanas se dever ao facto de que a AMP tem maior área de vilas pescatórias e população rural. [29] O modelo demonstrou ainda que, se todos os municípios tivessem a mesma população em relação às covariáveis do estudo, na AMP, a Póvoa de Varzim apresentaria a maior incidência média de TB; este município é caracterizado essencialmente pela actividade piscatória e agrícola. [35] Em contraste, Vila Nova de Gaia revelou ser o município com menor incidência média de TB, embora, sengundo dados do Census 2011, seja o maior concelho da área metropolitana (21% da área total da AMP), com o maior número populacional (aproximadamente 300000 habitantes) e com uma área rural significativa (18.8% da sua área total). [37] De facto, desde 2004, Vila Nova de Gaia implementou um programa de rastreio de contactos e de grupos de risco, desenvolvido por profissionais de saúde pública e médicos de família, com avaliação domiciliar e no local de trabalho de contactos de pacientes ativos de TB. Esta estratégia permitiu não só identificar os potenciais contactos em risco, como aumentar a adesão à triagem de TB de 67.6% para 87.3% dos contactos identificados e aumentar a taxa de conclusão do tratamento de 83% para 96%. [14]

A limitação principal deste trabalho, que na realidade seria inerente a qualquer estudo baseado nesta BD, foi o tamanho amostral. Com um total de 175 observações referentes a 35 municípios, não seria razoável construir um modelo com um elevado número de parâmetros. Assim, restringimo-nos a estruturas fixas e aleatórias com poucos parâmetros, assim como a estruturas de correlação e de heterocedasticidade pouco complexas. Outra limitação deste trabalho, resultante da dificuldade verificada na recolha de informação para a construção da BD, foi a falta de variáveis explicativas. Para além de existirem poucas plataformas de recolha de dados, a informação disponível nem sempre era referente aos anos pretendidos, pelo que houve variáveis

[1] Verdadeiras colmeias humanas, caracterizadas por um corredor aberto e estreiro a separar duas fileiras de casas muito pequenas, mal ventiladas e mal iluminadas, muitas delas com apenas uma divisão, com uma única casa de banho a ser compartilhada entre todos.

de interesse que não foram estudadas. Desta forma, como trabalho futuro, propõe-se não só a inclusão de variáveis que possam ser relevantes para o estudo (como variáveis que descrevam fatores sociais, nível de educação, alcoolismo, condições de habitação, sobrelotação, etc), assim como, se possível, a mudança da unidade de análise para o nível de freguesia, aumentando o tamanho da amostra e permitindo a inclusão de mais interações no modelo. Também se propõe a realização de estudos semelhantes em outros países europeus.

As principais conclusões decorrentes desta dissertação foram incluídas num artigo científico da área das doenças infeciosas, submetido para publicação. [16]

Bibliografia

[1] Akaike H. (1974). *A new look at the statistical model identification.* IEEE Transactions on Automatic Control 19, 716-723.

[2] Apolinário D., Ribeiro A.I., Krainski E., Sousa P., Abranches M. & Duarte R. (2017). *Tubercolosis inequelities and socio-economic deprivation in Portugal.* Int J Tuberc Lung Dis 21(7), 784-789.

[3] Box G., Jenkins G. & Reinsel G. (1994). *Time series Analysis: Forecasting and Control.* Holden-Day, San Francisco, 3 edition.

[4] Cabral M.S. & Gonçalves M.H. (2011). *Análise de Dados Longitudinais.* Sociedade Portuguesa de Estatística.

[5] Couceiro L., Santana P. & Nunes C. (2011). *Pulmonary tuberculosis and risk factors in Portugal: a spatial analysis.* Int J Tuberc Lung Dis 15, 1445-1430.

[6] Cressie N. (1993). *Statistical for Spatial Data.* Wiley, New York.

[7] Davidian M. & Giltinian D.M. (1995). *Nonlinear models for repeated measurement data.* Chapman & Hall, London.

[8] DeGruttola V., Lange N. & Dafni U. (1991). *Modeling the progression of hiv infection.* Journal of the American Statistical Association 86, 569-577.

[9] Dempster A., Laird N. & Rubin D. (1977). *Maximum likelihood for incomplete data via em algorithm.* Journal of the Royal Statistical Society, Series B 39, 1-38.

[10] Dias M., Gaio R., Sousa P., Abranches M., Gomes M., Oliveira O., Correia-Neve M., Ferreira E. & Duarte R. (2017). *Tuberculosis among the homeless: should we change the strategy?.* Int J Tuberc Lung Dis 21(3), 327-332.

[11] Direção-Geral da Saúde, Ministério da Saúde, Portugal. *Programa Nacional de Luta contra a Tuberculose. Ponto de situação epidemiológica e de desempenho. Dia Mundial da Tuberculose.* Portugal: Ministério da Saúde, 2013.

[12] Direção-Geral da Saúde, Direção de Serviços de Informação e Análise, Ministério da Saúde, Portugal. *Programa Nacional para a Infeção VIH, SIDA e Tuberculose 2017.* Lisboa, Portugal: Ministério da Saúde, 2017.

[13] Direção-Geral da Saúde, Direção de Serviços de Informação e Análise, Ministério da Saúde, Portugal. *Programa Nacional para a Infeção VIH, SIDA e Tuberculose 2016*. Lisboa, Portugal: Ministério da Saúde, 2016.

[14] Duarte R., Neto M., Carvalho A. & Barros H. (2012). *Improving tuberculosis contact tracing: the role of evaluations in the home and workplace*. Int J Tuberc Lung Dis 16(1), 55-59.

[15] European Centre for Disease Prevention and Control (ECDC). *Tb and HIV co-infection in the EU/EEA. ECDC, 2017*. Disponível em: `https://ecdc.europa.eu/en/publications-data/tb-and-hiv-co-infection-eueea`

[16] Felgueiras M., Cerqueira S., Gaio R., Felgueiras O. & Duarte R. (2017). *A comparative study of the tuberculosis incidence between the two main portuguese metropolian areas*. Artigo submetido para publicação.

[17] Fernandes P. *A cidade do Porto na 1ª metade do século XIX: população e urbanismo*. Centro de Estudos da População, Economia e Sociedade (CEPESE), 1996. Disponível em: `http://www.cepesepublicacoes.pt/portal/pt/obras/populacao-e-sociedade/revista-populacao-e-sociedade-no-2/a-cidade-do-porto-na-1-a-metade-do-seculo-xix-populacao-e-urbanismo`

[18] French C., Kruijshaar M., Jones J. & Abubakar I. (2009). *The influence of socio- economic deprivation on tuberculosis treatment delays in England, 2000-2005*. Epidemiol Infect 137, 591-596.

[19] Fitzmaurice G., Laird N. & Lipsitz S. (1994). *Analyzing incomplete longitudinal binary responses: A likelihood-based approach*. Biometrics 50, 601-612.

[20] Fitzmaurice G., Laird N. & Ware J. (2004). *Applied Longitudinal Analysis*. Wiley, New York.

[21] Franco I., Sousa P., Gomes M., Oliveira A., Gaio A.R. & Duarte R. (2016). *Social profile of the highest tuberculosis incidence areas in Portugal*. Rev Port Pneumol 22(1), 50-56.

[22] Gurka M.J. (2006). *Selecting the Best Linear Mixed Model under REML*. The American Statistician 60(1), 19-26.

[23] Hanway A., Comiskey C., Tobin K. & O'Toole R.F. (2016). *Relating annual migration from high tuberculosis burden country of origin to changes in foreign-born tuberculosis notification rates in low-medium incidence European countries*. Tubercolosis 101, 67-74.

[24] Harville D. (1977). *Maximum likelihhod approaches to variance component estimation and to related problems*. Journal of American Statistical Association 72, 320-338.

[25] Hollo V., Kotila S.M., Kodmor C., Zucs P. & van der Werf M.J. (2016). *The effect of migration within the European Union/European Economic Area on the distribution of tuberculosis, 2007 to 2013*. Euro Surveill 21(12), pii=30171.

[26] Kodmor C., Zucs P. & van der Werf M.J. (2016). *Migration-related tuberculosis: epidemiology and characteristics of tuberculosis cases originating outside the European Union and European Economic Area, 2007 to 2013*. Euro Surveill 21(12), pii=30164.

[27] Lehmann E.L. (1986). *Testing Statistical Hypotheses*. Wiley, New York.

[28] Lopez de Fede A., Stewart J., Harris M. & Mayfield-Smith K. (2008). *Tuberculosis in socio-economically deprived neighborhoods: missed opportunities for prevention*. Int J Tuberc Lung Dis 12(12), 1425-1430.

[29] Matos F. *A habitação no grande Porto: uma perspectiva geográfica da evolução do mercado e da qualidade habitacional desde finais do séc. XIX até final do milénio*. Universidade do Porto, Faculdade de Letras, 2001.

[30] Patterson H. & Thompson R. (1971). *Recovery of inter-block information when block sizes are unequal*. Biometrika 58, 545-554.

[31] Pimpim L., Drumright L.N., Kruijshaar M.E., Abubakar I., Rice B., Delpech V., Hollo V., Amato-Gauci A., Manissero D. & Kodmon C. (2011). *Tubercolosis and HIV co-infection in European Union and European Economic Area countries*. Eur Respir J 38, 1382-1392.

[32] Pinheiro J. & Bates D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, New York.

[33] Ponticiello A., Sturkenboom M., Simonetti A., Ortolani R., Malerba M. & Sanduzzi A. (2005). *Deprivation, immigration and tuberculosis incidence in Naples, 1996-2000*. Eur J Epidemiol 20, 729-734.

[34] R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria, 2016. Disponível em: `https://www.R-project.org/`

[35] Rede Social Póvoa do Varzim. *Diagnóstico Social: Concelho da Póvoa de Varzim, 2006*. Disponível em: `http://www.cm-pvarzim.pt/areas-de-atividade/coesao-social/rede-social/documentos-produzidos-no-claspv-1/diagnostico-social-do-concelho-da-povoa-de-varzim`

[36] Schwarz (1978). *Estimating the dimension of a model*. The Annals of Statistics 6, 461-464.

[37] Silva V. & Graça P. (2006) . *Plano Director Municipal de Gaia (PDM). Relatório 2.6 – Caracterização Biofísica*. Município de Vila Nova de Gaia, Agosto de 2006.

[38] Stram D. & Lee J. (1994). *Variance components testing in the longitudinal mixed effects model*. Biometrics 50, 1171-1177.

[39] Taylor B.M. & Nunes C. (2016). *Modelling the time to detection of urban tuberculosis in two big cities in Portugal: a spatial survival analysis*. Int J Tuberc Lung Dis 20(9), 1219-1225.

[40] Thisted R. (1988). *Elements of Statistical Computing*. Chapman & Hall, London.

[41] van Hest N.A., Aldridge R.W., de Vries G., et al. (2014). *Tuberculosis control in big ciries and urban risk groups in the European Union: a concensus statement*. Euro Surveill; 19(9):pii=20728.

[42] Verbeke G. & Molenberghs G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

[43] West B.T., Welch K.B. & Galecki A.T. (2007). *Linear mixed models: A practical guide using statistical software*. Chapman & Hall/CRC, Boca Raton.

[44] World Health Organization (WHO). *Global tuberculosis report 2016*. Disponível em: `http://www.who.int/tb/publications/global_report/en/`

[45] World Health Organization (WHO). *Tubercolosis surveillance and monitoring in Europe 2016*. Disponível em: `https://ecdc.europa.eu/sites/portal/files/media/en/publications/Publications/ecdc-tuberculosis-surveillance-monitoring-Europe-2016.pdf`

[46] Zuur A., Ieno E., Walker N., Saveliev A. & Smith G. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York.